# Functional characterization of genetic alterations in cancer

# Functional characterization of genetic alterations in cancer

A dissertation presented

by

Eejung Kim

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

April 2016

Dissertation Advisor: Dr. William Hahn Eejung Kim

# Functional characterization of genetic alterations in cancer

## Abstract

The comprehensive identification of genetic alterations is critical to understanding the pathophysiology of cancer. Recent advances in sequencing technology have enabled the detailed description of cancer genomes. However, to translate these findings into a deeper understanding of cancer biology, analyzing the functional impact of cancer-associated genetic aberration is essential. Here I investigate how to accelerate the functional characterization of two classes of genetic alterations, point mutations and amplifications.

The wide spectrum of point mutations that arise in cancer makes them challenging to study comprehensively. I have developed a scalable systematic method to experimentally infer the functional impact of cancer-associated gene variants. I performed pooled *in vivo* tumor formation assays and gene expression profiling using 474 mutant alleles curated from 5,338 human tumors. I identified 12 transforming alleles including two in genes (*PIK3CB*, *POT1*) that have not been previously shown to be tumorigenic. One rare *KRAS* allele, D33E, displayed tumorigenicity and constitutive activation of RAS effector pathways. By correlating gene expression changes induced upon expression of wild type and mutant alleles, I could infer the activity of specific alleles. These approaches enable the interrogation of cancer-associated alleles at scale and demonstrate that rare alleles may be functionally important.

Frequently amplified regions in cancer often harbor oncogenic drivers. However, identifying the driver gene among many other amplified genes is challenging. In high-grade serous ovarian cancer (HGSOC), 1,825 genes are amplified across 63 amplicons. We employed systematic loss-of-function RNAi data to identify amplified genes that were essential

in the ovarian lineage. We identified 50 amplified and essential genes and validated *FRS2*, an adaptor protein in FGFR pathway. *FRS2*-amplified cancer cell lines were dependent on *FRS2* expression and *FRS2* overexpression in immortalized cell lines was sufficient to promote anchorage independent growth and tumorigenesis in nude mice. This approach demonstrates that intersecting structural genomics with functional genomics can facilitate the discovery of driver genes in recurrently amplified regions. Collectively, the methods I present here provide a framework to study point mutations and amplifications to accelerate the interpretation of the cancer genome.

# Table of Contents

# Acknowledgements

I have been very fortunate to receive support, guidance and kindness from many people who have contributed greatly to my scientific growth. Bill Hahn, my thesis advisor, has been an inspirational mentor who encouraged me to pursue my scientific interest while making sure that I have tangible results. He taught me to be an independent scientist by providing constant support and learning opportunities. Thanks to him, I feel confident that I can pursue any scientific questions and arrive at an answer with competence, fervor and enthusiasm. I am also immensely grateful to Seong-Jin Kim, who introduced me to the excitement of research while I was in medical school and encouraged me to study abroad to expand my research experience.

I am honored to have met and worked with many inspiring scientists over the years. Alicia Zhou and Rhine Shen are incredible scientists who were my science moms. They raised me through my scientific infancy by answering thousands of “what’s that?” and “but why?” questions as well as teaching me experimental techniques and how to plan, troubleshoot and interpret experiments. Sefi Rosenbluh and Deepak Nijhawan transformed me from a science hobbyist to a true disciple through their infectious enthusiasm for science.

This dissertation would not have been possible without Andy Aguirre, who sustained me through difficult times by encouraging me to believe in myself and providing valuable advices on research and career. I consider myself the most fortunate to have had Andy’s mentorship during graduate school. Equally indispensible was the support and friendship from Belinda Wang, who has been incredibly generous with her time and advice. She also taught me how to write coherently and professionally. I will be forever grateful for all her last minute editing.

I would like to thank my collaborators, Jesse Boehm, Gad Getz, Nina Ilic, Atanas Kamburov, Yash Shreshtha, Lihua Zou, John Doench, Cong Zhu, David Root, Leo Luo and

Tony Cheung who contributed greatly to the two projects I conducted in graduate school. I have learned a lot from their expertise and expansive knowledge.

Past and present member of the Hahn lab, Atish Choudhury, Andy Hong, Yaara Zwang, Andrew Giacomelli, Nikki Spardy, David Takeda and Elsa Krall have greatly enriched my graduate school experience by providing inspirational conversation and valuable research and career advice.

I would also like to thank members of the Broad Cancer Program who have empowered me to become a better scientist. Eran Hodis taught me programming skills and encouraged me to see scientific questions with rigor and quantitative perspective. Zuzana Tothova has provided invaluable advice on science and career development. I am greatly indebted to her. I appreciate Angela Brooks and Alice Berger for generous advice and insightful discussions.

I am grateful for my Dissertation Advisor Committee members, Matthew Meyerson, Benjamin Ebert and Rameen Beroukhim for their scientific acumen and constructive criticism. I would also like to thank my Defense Committee members, Cathy Wu, Kevin Haigis and Carla Mattos for their generosity and for taking the time to read my dissertation.

Finally, I would like to thank my family and friends for their unwavering support and love. My late mother taught me the love of learning for the sake of learning by constantly teaching herself new things. My father has been my best friend, therapist, source of confidence and pride throughout my life. He instilled in me the love of nature through numerous backpacking, hiking and camping trips and taught me how to think logically and express myself coherently by engaging me in debates and discussions for hours everyday.

# Chapter 1

# Introduction: Exploring genetic alterations in cancer

## 1.1 Cataloguing genetic alterations in cancer

The comprehensive description of genetic alterations in cancer has been a major goal of cancer research, with the expectation that identifying these aberrations would elucidate the molecular basis of cancer and nominate potential therapeutic targets (1-3). This expectation is based on prior triumphs, such as the discovery of the *BCR-ABL* translocation (4,5), *HER2* amplification (6) and *BRAF*$^{V600E}$ point mutation (7), which led to the development of efficacious targeted cancer therapeutics, imatinib, tratuzumab and vemurafenib (8-10).

These success stories were preceded by decades of relentless searching for cancer causing genes. The hunt for cancer causing genes evolved in parallel with advancements in detection methods. As the detection technology became more sophisticated, many more genetic aberrations were discovered. The concept of cancer as a disease that evolves from somatic genetic alterations originated from the study of cancer-inducing retroviruses in animal models. These retroviruses were found to contain oncogenic genes such as Src and Ras (11,12). Using transforming retroviruses' sequences as probes, homologous genes in human genome were discovered (13). Progress in gene transfer technology, such as transfection and retroviral delivery, enabled the detection of DNA fragments extracted from cancer cells that were capable of transforming non-cancer cells; these methods facilitated the discovery of an oncogenic point mutation in *HRAS,* G12V (14,15).

Conventional cytogenetic analysis using light microscope facilitated the discovery of abnormal chromosomal rearrangements (4,5), and genetic amplifications (16). Advancements in molecular cytogenetic technologies, such as fluorescent *in situ* hybridization (FISH) and array comparative genomic hybridization (aCGH), have greatly improved the resolution and accuracy of detecting chromosomal aberrations. To date, hundreds of translocations and copy number alterations have been identified (17).

The development of high throughput methods, such as aCGH and gene expression microarrays, allowed discovery of novel cancer associated genetic alterations. Indeed, analyzing gene expression outliers by microarray identified *TMPRSS2-ERG* translocation in prostate cancer (18). In addition, systematic loss-of-function and gain-of-function screening methods have facilitated the identification of genes that contribute to the development and progression of human cancers. RNAi screening enabled the identification of several tumor suppressor genes (19). A complementary DNA library screen identified *EML4-ALK* fusion as an oncogenic driver in lung cancer (20), and a kinome overexpression screen identified *IKBKE* as a breast cancer oncogene (21).

After the completion of the Human Genome Project (22), targeted PCR followed by Sanger sequencing became the predominant method to discover novel oncogenic point mutations in candidate genes. Several *BRAF* (7), *ERBB2* (23), *PIK3CA* (24,25), *JAK2* (26) and *AKT1* (27) point mutations were discovered in this manner. The number of genes studied increased rapidly, and by 2007, such study of exome-scale was performed in breast and colorectal cancers (28). These unbiased whole exome studies empowered the discovery of unexpected, novel cancer associated genes, such as *IDH1* (29).

Even though the methods described above allowed discovery of many oncogenes and tumor suppressor genes, they had two inherent problems: researchers needed some *a priori* knowledge of what they were looking for, and most of these methods did not scale well due to their laborious nature and/or high costs. Introduction of the next-generation sequencing in the mid 2000s, effectively addressed these issues by driving down the cost of sequencing to orders of magnitude lower than that of Sanger sequencing, while maintaining high accuracy (30).

As sequencing the whole exome and genome became feasible, cancer researchers quickly adapted next-generation sequencing technology to comprehensively describe genetic, epigenetic, and transcriptomic alterations in cancers. In the past ten years, the detailed description of the mutational landscape in many types of cancers was accomplished (31,32).

International collaborative efforts, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), have greatly expedited this process by each characterizing more than 10,000 tumors of 20 cancer types and more than 25,000 tumors of 50 types, respectively (1,33). As the cost of sequencing plummeted, it became financially feasible for individual investigators to sequence hundreds of matched tumor and normal tissue pairs. The exponential increase in the number of tumors sequenced facilitated discovery of not only novel cancer-associated genes in previously known pathways, but also entirely new classes of genes involved in cellular processes such as epigenetic modifications, splicing regulations and protein homeostasis (31). These new findings have introduced exciting new fields of investigation to the cancer research community that may hold keys to novel mechanistic understanding and therapeutic developments.

The bulk of sequenced tumors to date have been analyzed by whole-exome sequencing, which is limited to sequencing the protein coding part of the genome. However, the coding sequence in the human genome accounts for only 1-2% of the total sequence; the rest of the genome may hold important information in understanding cancer biology (31). With decreasing sequencing costs, whole-genome sequencing is increasingly applied to study genetic alterations in cancer, enabling the identification of alterations that cannot be captured in exome sequencing. These include point mutations in noncoding regions such as regulatory elements and long noncoding RNAs (34), as well as complex genetic rearrangements including multiple translocations, chromoplexy (35), and chromothrispsis (36,37). Point mutations in regulatory elements may play a significant role in cancer. Recently, point mutations in the telomerase reverse transcriptase (*TERT)* promoter were discovered in melanoma and other cancers, and were shown to reactivate telomerase activity in cancer cells (38-40). Whole genome analysis of hundreds of human tumors has identified multiple recurrent mutations in the upstream region of genes such as *PEKHS1*, *WDR74* and *SDHD,* which may play a role in cancer initiation and progression (41).

Along with mutations in noncoding regions, the functional impact of synonymous point mutations in coding region is increasingly investigated. Synonymous variants have been shown to affect protein expression, conformation and function through distinct mRNA processing, mRNA secondary structure, and post-transcriptional regulations (42). Synonymous mutations in *BCL2L12* can change miRNA-mediated regulation, which results in increased mRNA and protein levels in malignant melanoma (43). A recent survey of exome sequencing data from more than 3,000 tumors showed that some synonymous variants are indeed under positive selection, and these variants are concentrated in oncogenes and 3'UTRs, affecting splice sites and mRNA expression levels (44). These discoveries of functional variants in noncoding regions and synonymous mutations will increase as more cancer genomes are sequenced and will continue to enrich the mechanistic understanding of cancer initiation and progression.

In this thesis, I focus on the functional characterization of two classes of genetic alterations, non-synonymous point mutations and focal amplifications. In the following section, I will explore current challenges of investigating these alterations.

## 1.2 Challenges of translating the cancer genome

### 1.2.1 Interpreting non-synonymous point mutations

These massive cancer genome characterization efforts described above have yielded valuable insights on cancer biology, but the amount of data generated has already far surpassed our ability to analyze and interpret. As of August 2014, The Catalogue of Somatic Mutation in Cancer (COSMIC), a comprehensive cancer mutation repository, has described over two million coding mutations curated from more than one million tumors, published in about 20,000 research papers (45). When only whole exome or whole genome sequencing data were considered to remove selection bias for well-known cancer genes, 17,457 tumors with at least one non-synonymous point mutation were curated in COSMIC as of April 2016. These tumors

harbored 1,706,408 unique non-synonymous point mutations and 1,533,746 of them (~90%) were observed only once (**Figure 1-1**). It is likely that the vast majority of these millions of point mutations are "passenger" mutations, mutations that do not confer selective advantage to tumor cells (46). On the contrary, when the number of times each allele was mutated was examined, there were many alleles with recurrent mutations; for example, 4,211 unique alleles from 1,508 genes were observed more than 10 times each, implying that these recurrently mutated alleles are likely to be functionally important in cancer (**Figure 1-1**). However, the majority of these alleles and genes have not yet been studied in the context of carcinogenesis. The sheer number of non-synonymous point mutations found in cancer genome makes in-depth study of every allele prohibitively resource demanding.

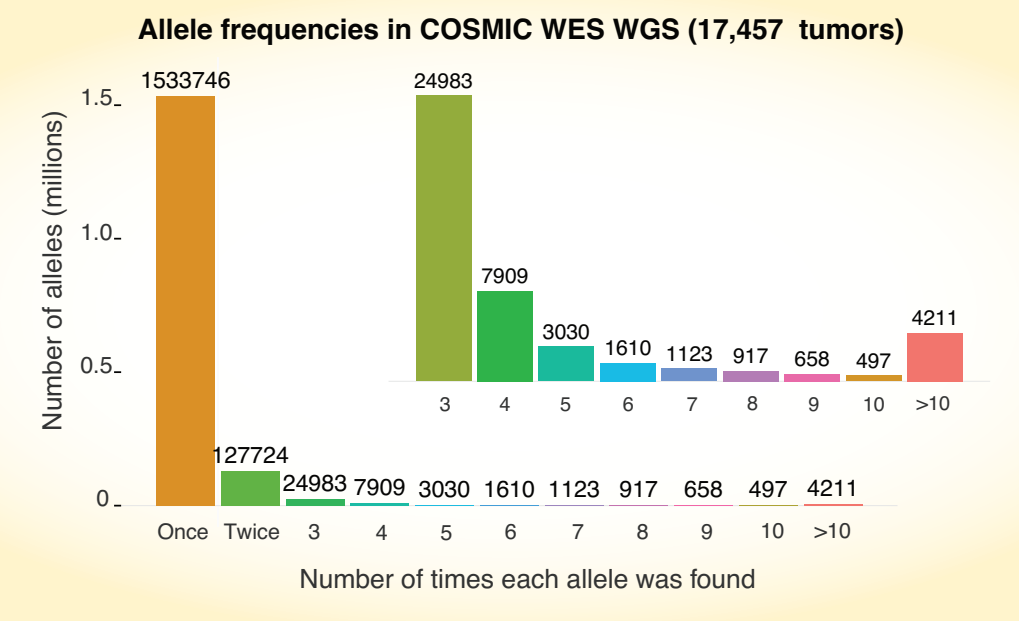



**Figure 1-1. Allele frequencies in COSMIC whole-exome and whole-genome sequencing data.**

More than 90% of non-synonymous point mutations reported were observed only once. Alleles that were observed three or more times were shown in the inset for easier comparison. 24,983 unique alleles were observed three times. 4,211 alleles were observed more than ten times.

Furthermore, the rarity of specific alleles makes the statistical inference of importance based on the incidence impossible to perform. Currently, significantly mutated genes in cancer genome sequencing research are calculated by accounting for gene-specific background mutation rates, which are determined by considering variables such as gene length, nucleotide composition, distance to telomeres and the centromere, mutation rates in the intron sequences, gene expression level, and replication timing (47-50). These gene-level significance calculations have reduced false positive reports of significantly mutated genes over the years (51-53). While variables affecting the gene-specific background mutation rates are being actively investigated, there is currently no consensus on how to account for the allele-specific background mutation rate (Michael Lawrence, personal communication). Furthermore, the rarity of the majority of cancer-associated non-synonymous point mutations makes recurrence-based prioritization challenging. Even in well-characterized oncogenes and tumor suppressor genes, the majority of rare alleles have not yet been studied (54). To predict the functional impact of these mutations, parameters such as evolutionary conservation, biochemical properties of amino acids and existence of the allele within a known functional domain have been used to predict the functional consequences of the amino acid substitution (50,54,55). Popular *in silico* methods utilizing these parameters include Polyphen2 (56), Mutation Assessor (57), CHASM (58), VEST (59) and SIFT (60). Even though these algorithms can provide additional information on the functional impact of point mutation, all these methods suffer from two major problems: one is inadequate sensitivity and specificity, which were reported to range from 40 to more than 90% (50), and the other is limited prediction of functional consequence. These methods assess whether the point mutations would affect the function of the protein, but not whether the effect would be gain- or loss- or switch- of functions (61).

The challenge of abundance and rarity in characterizing cancer-associated point mutation is evident even in the most well characterized oncogenes and tumor suppressors. For example, *KRAS*, one of the most commonly mutated oncogenes, has a distinct hot spot

mutation site; however, there are also many alleles, including 43 that have been observed only once, that have not been studied (**Figure 1-2**). As *KRAS* mutation status is already used in directing monoclonal antibody therapy against EGFR in metastatic colorectal cancer (62,63), it is important to understand whether these alleles affect KRAS function.

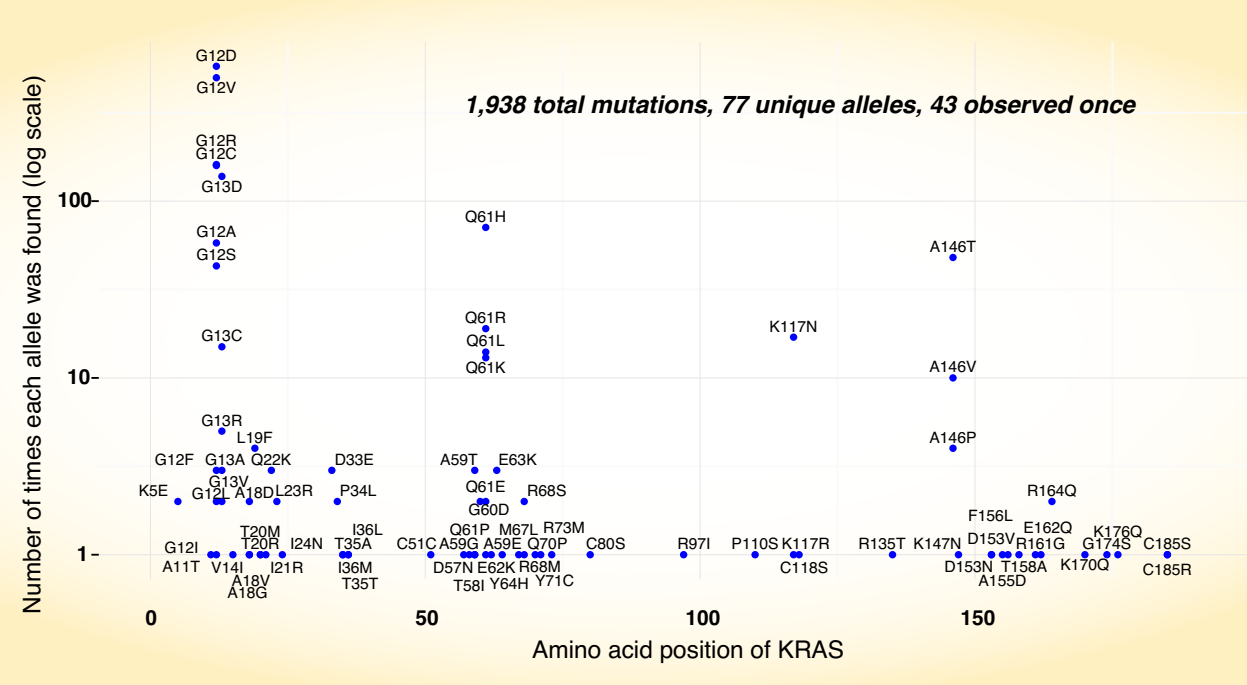



**Figure 1-2. Allele frequencies in *KRAS* across the gene length.**

X-axis shows the amino acid position of KRAS. Y-axis shows the incidence of certain alleles. 1,938 mutations were found, which belong to 77 unique alleles. 43 of these were found only once. All alleles were labeled.

The abundance and rarity problem applies to both oncogenes and tumor suppressors, but the magnitude of the problem can be greater in tumor suppressor genes because unlike oncogenes, which tend to have a concentrated region of higher mutation burden, tumor suppressors tend to have mutations scattered across the length of the gene (32). For example, when COSMIC data were analyzed, *PTEN*, a well-known tumor suppressor gene, was found to have hundreds of non-synonymous mutations that were observed only once (**Figure 1-3**). Out

of 235 unique mutated alleles in *PTEN*, 175 were observed only once. Differentiating loss-of-function mutations from the passenger mutations based on incidence is not possible. *PTEN* null status, along with *PIK3CA* mutation status, is used for determining eligibility for enrolling in clinical trials for agents targeting PI3K/AKT/mTOR pathways (64). In the case of truncating, nonsense mutations, determining PTEN status is straightforward; however, the case of point mutation needs further investigation.

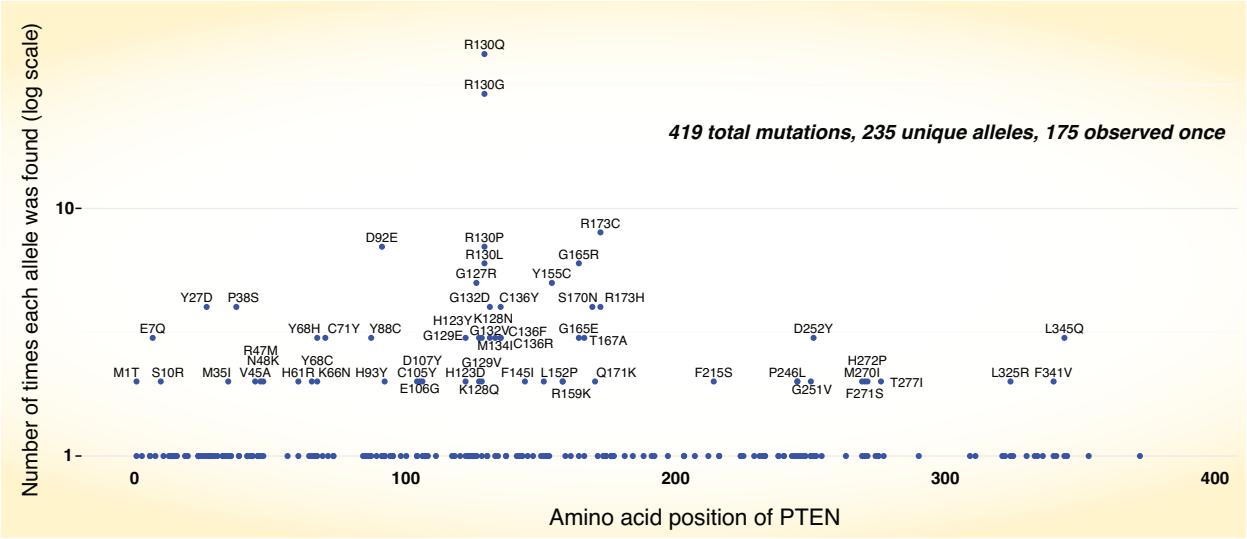



**Figure 1-3. Allele frequencies in *PTEN* across the gene length.**

X-axis shows the amino acid position of PTEN. Y-axis shows the incidence of certain alleles. 419 mutations were found, which belong to 235 unique alleles. 175 of these were found only once. Mutated alleles with two or higher incidence were labeled.

In summary, the abundance of observed mutations, rarity of certain alleles, and insufficient predictive value of computational algorithms require the development of new methodologies to address the challenge of characterizing non-synonymous point mutation in cancers. Ideally, these methods should provide functional information and achieve scalability (**Figure 1-4**).

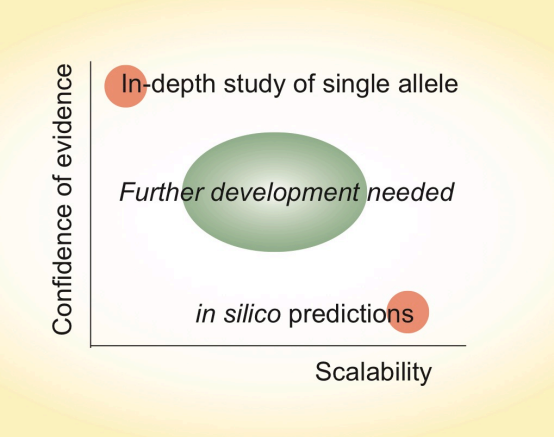



**Figure 1-4. New methods are needed in characterizing point mutations in cancer.**

### 1.2.2 Identifying the driver gene in somatic copy number alterations

The number of point mutations that have been identified in cancer genomes is enormous. Equally daunting is the number of aberrations found in other categories of genetic alterations, such as amplification. A recent analysis of 4,934 tumors from TCGA dataset showed that there are 152 regions of recurrent somatic copy number alterations (SCNA), 102 of which did not harbor known oncogenes or tumor suppressors genes (65).

Since *MYC* was identified as an amplified oncogene (66,67), many other amplified oncogenes were discovered; discovery of *ERBB2* (*HER2*) amplification in breast cancer (68) led to development of an effective targeted therapeutics (10). To credential an amplified gene as a bona fide oncogene, several criteria need to be met. These criteria include: evidence of recurrent amplification containing the candidate gene, correlation between amplification and overexpression of the gene, biological and/or clinical adverse outcomes that are associated with overexpression of the gene, and essentiality of the gene in cancer cells harboring the amplification (69). The central challenge in identifying the driver genes in amplification regions is twofold: as in the case of non-synonymous point mutation, “passenger” amplifications can also be fixed in the cell population and “driver” amplifications can harbor thousands of genes (70).

Taking into account the frequency and amplitude of amplifications across many tumor samples helped distinguish “driver” amplifications from “passenger” amplifications. However, accurately estimating background SCNA rate is still under active investigation (70,71). Identifying the driver gene among many other amplified genes is another major challenge in characterizing putative driver amplifications (32,72). When the number of genes in the amplicon is small or when there is already a cancer-associated gene within the amplicon, the identification of driver gene can be relatively straightforward by using candidate approach (73-75). However, when many genes are located within the same amplicon, identifying the driver gene is more complicated.

Employing systematic loss- and gain-of-function genetic perturbation can reduce the search space by providing orthogonal filters. For example, the recurrent amplification on chromosome 22q11.21 was interrogated by utilizing RNAi screening to nominate *CRKL* as the driver gene (76). Subsequent in-depth investigation of *CRKL* elucidated the mechanism of *CRKL*-mediated transformation and resistance to EGFR inhibitor therapy (77,78). An alternative approach in identifying the driver gene of an amplified region is to query every gene in the amplicon or set of amplified genes in specific types of cancer. 124 amplified genes in hepatocellular carcinoma were screened by cDNA overexpression, and *FRF19* and other 17 genes were nominated as driver oncogenes (79). Hagerstrand and colleagues systematically interrogated 20 genes within the amplicon on chromosome 3q26 by shRNA knock-down and open reading frame (ORF) overexpression experiments and identified *TLOC1* and *SKIL* to be driver genes (80). A genome scale ORF overexpression screen identified *MECP2* as a potential oncogene (81). These systematic functional genomics screens have greatly increased the power to detect the driver genes in the amplified regions. However, many recurrently amplified genomic regions remain poorly studied. In this thesis, I integrate structural and functional genomics in high-grade serous ovarian cancer cell lines to identify driver genes.

## 1.3 Nomenclature of cancer-associated somatic alterations

The nomenclature of somatic mutations in cancer is still evolving and it is important to specify definitions to facilitate precise description and discussion. In this thesis, following terms will be used as described here. “Driver gene/mutation/amplification” refers to a gene/mutation/amplification that facilitates the proliferative advantage and consequent positive selection of cells harboring that alteration; “passenger gene/mutation/amplification” does not confer such advantage (32,46). “Functional mutation” or “functional variant” (used interchangeably) refers to a mutation that affects the function of a protein and includes gain, loss and switch of function variants (61). “Neutral mutation/variant” indicates a mutation that does not change the molecular function of the protein. A functional mutation is not necessarily a driver mutation for two reasons: mutations can be biochemically functional but biologically inert, and biologically functional mutations may not always confer a selective advantage in the native environment of the tumor cells (55). “Functional mutation” is further divided into “gain-of-function (GOF) mutation,” “loss-of-function (LOF) mutation” and “change-of-function (COF) mutation” based on the functional consequence of such mutation, when the effect is compared to that of the wild type (**Figure 1-5A**). Precisely defining the term “functional” is challenging due to two reasons: we currently do not know all the functions of proteins and the functional impact of a variant may not be binary.

For instance, *PTEN*, a well-known tumor suppressor, plays multiple functions. It is a dual function protein and phospholipid phosphatase that homodimerizes to increase its activity (82). Many more functions of *PTEN* are currently known, but in a simplified three-function model, “loss-of-function mutation” technically can be a variant that has a decreased activity in any of these three functions (**Figure 1-5B**). Even though the lipid phosphatase function of PTEN is known to be important in driving tumorigenesis (83,84), studies on the relevance of other function on tumorigenesis are still undergoing; variants affecting its localization to the plasma

membrane have recently found to be important for PTEN to impart its lipid phosphatase function (85). Pinpointing cancer relevant functions among many known as well as yet-undiscovered functions of a single protein is challenging.

Another challenge of designating non-synonymous point mutation as a “functional variant” is that the functional consequences of the amino acid substitution may not always be binary. It is common for variants to partially lose its native function or to confer slight activation of cancer related molecular pathways. When the lipid phosphatase function of 40 different PTEN alleles was interrogated, many alleles exhibited partial LOF (83). When different KRAS alleles were tested for downstream pathway activation, they showed gradients of differential activation (86,87). These cases demonstrate the difficulty of drawing the cutoff to call variants “functional” (**Figure 1-5C**). Variants with subtle change in their function, in combination with other genetic alterations, may be sufficient to drive tumorigenesis, as in the case with other complex diseases. In this thesis, I consider alleles with even subtle phenotypic changes to be functional.

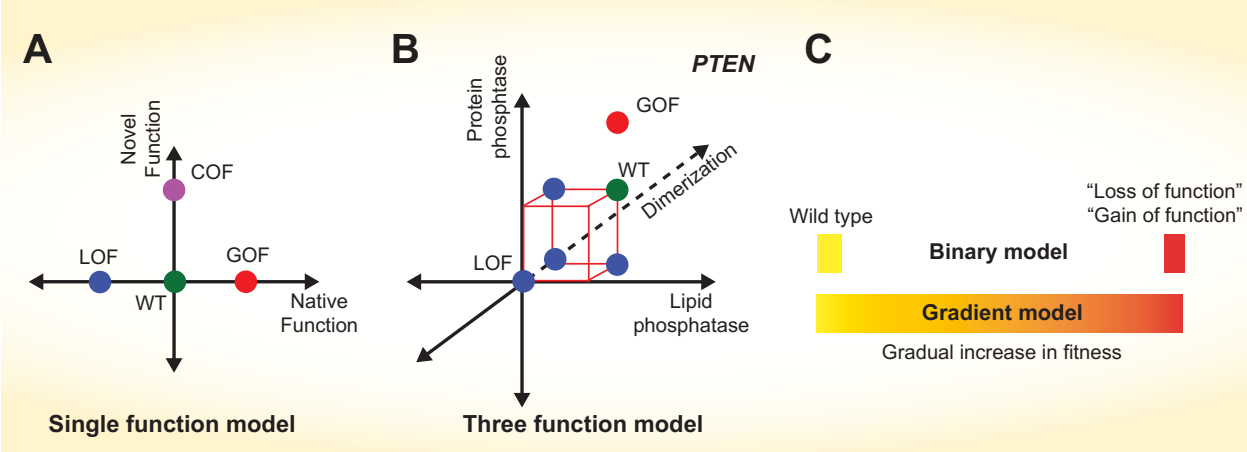



**Figure 1-5. Interpreting functional impacts of non-synonymous point mutations.**

(**A**) In the single function model, a variant with enhanced native function is called gain-of-function (GOF) variant. A variant with decreased native function is called loss-of-function (LOF) function. A variant that has acquired new function is called change-of-function (COF) variant.

**Figure 1-5. (Continued).**

(**B**) In the three-function model, such designation may not be straightforward. Variants with decrease in any of the native functions of the wild type protein should be designated as LOF variant.

(**C**) In many cases, functional impact of allele is not binary but more gradual. Determining the cutoff for functional allele is challenging.

## 1.4 Experimental methods to interrogate genetic alterations in cancer

### 1.4.1 In-depth interrogation of a single alteration

Most of the mechanistic understanding of cancer biology is derived from decades of contributions from many researchers studying one genetic alteration in cancer. These efforts involve cloning the mutant gene and comparing the effect of the mutant ORF with the wild type ORF in various phenotypic assays, such as biochemical properties, molecular pathway activation/inhibition, anchorage independent growth, *in vivo* tumorigenesis, or invasion and migration (88). One of the most convincing methods to determine whether a genetic alteration contributes to tumorigenesis is to recreate the same genetic lesion in animal models and observe latent tumorigenic phenotype (89,90). Since the advent of transgenic, knockout and knock-in mice technology, many canonical oncogenes and tumor suppressors have been identified and validated using this method (91,92). With the introduction of inducible and tissue-specific transgenic and knock-in models, more sophisticated control over gene expression has been achieved, enabling the assessment of dependency on particular oncogenic alleles for tumor maintenance and reversibility of tumor phenotype upon reactivation of tumor suppressors (93-96). Crossing mouse strains harboring different genetic alterations has also allowed the study of combinatorial effect of these alterations. Subtle effects of single genetic alterations may

be potentiated in the presence of additional alterations (97). These models have been invaluable not only to elucidate the causal effect of genetic alterations, but also to study the evolution of tumors, interaction of tumor cells with their microenvironments, and response to putative therapeutic agents (98). The wealth and depth of knowledge gained from in-depth studies of a single alteration have contributed immensely to our understanding of cancer and this method will undoubtedly continue to be the mainstay of studying novel oncogenic alleles. However, as this method of interrogating functional consequence typically requires years of effort and multiple researchers, it is not readily scalable.

### 1.4.2 Investigating many alleles of one gene with phenotypic assays

Once the function of a canonical allele in an oncogenes or tumor suppressor is studied in-depth, many alleles of the same gene can be studied with the phenotypic assay that measures the known function. For known cancer associated genes, examining the effect of multiple alleles can identify novel functional domains and novel interactions between proteins. Importantly, this type of research can be helpful in accurately annotating rare alleles in genes that are currently sequenced for clinical purposes. Genes such as *BRCA1/2*, *KRAS*, *BRAF* and *EGFR* are currently being sequenced in clinical setting to guide genetic counseling, clinical trial enrollment and therapeutic decisions (99-103). Notably, even in one of the best characterized tumor suppressors, *BRCA1*, whose LOF mutation underlies a hereditary cancer syndrome, 5 to 20% of the testing results currently report a variant of unknown clinical significance (VUS) (99), meaning that the functional consequences of the specific allele have not been characterized. As the cost of synthesizing oligonucleotide decreases, cloning multiple alleles of the same gene has become increasingly manageable (104). Recent advancements in cloning technology enable the efficient construction of expression libraries containing every single possible allele of a single gene (105). This approach may address the problem of abundance and rarity in known

cancer associated genes; however, it is unable to identify rare but novel cancer associated variants.

### 1.4.3 ORF gain-of-function screen

Screening putative oncogenes in large scale has facilitated discovery of novel oncogenes. When investigating focal amplification harboring multiple genes, systematically overexpressing individual genes and assessing the functional consequence can enable the identification of driver genes in the amplicon. The same approach can be used to delineate driver genes among many other overexpressed genes. Since the construction of human ORFeome collection, screening many more genes simultaneously, even at a near genome scale, has become feasible (106,107). However, current ORF collections mostly contain wild type ORFs, and few altered alleles are included. Constructing ORF collections including many alleles with genetic alterations found in cancer could be a powerful way to facilitate the study of these alleles at scale.

### 1.4.4 shRNA loss-of-function screen

One way to identify cancer specific essential genes is to remove the effect of the genes in both cancer and normal settings and to show the specific importance of the gene in cancer. In studying focal amplification, knocking down genes in the amplicon systematically can facilitate identification of amplified and essential genes. Genome scale shRNA loss-of-function screens can also identify genes important in specific lineages or in tumors with specific genetic alterations. Genes important for survival in cancer cells with ovarian lineages and genes in synthetic lethal relationship with *KRAS* activating mutation were discovered via this method (76,108). Off-target effect of the shRNA, shRNAs targeting mRNAs with incomplete sequence match has been well reported, but the mechanism is not completely understood (107,109,110).

To adjust for such effects, various computational as well as experimental methods have been developed. Measuring the consensual effects among multiple shRNAs (111), constructing complementary set of shRNAs that share seed sequences with the corresponding shRNA (112), and rescuing the effect of shRNA with matched ORF overexpression that are resistant to shRNA have all increased the specificity of interpreting shRNA screen data (109). Though imperfect, shRNA screening data have provided valuable insights to gene function in cancer.

### 1.4.5 Gene expression as a readout of functional impact

Since its introduction in the 1990s, gene expression profiling has been extensively tested as a measure for tumor type sub-classification, prediction of response to therapy, prognostic correlation and specific pathway activation (113). Recently, gene expression was proposed as a generalizable readout of cellular state that is achieved by genetic or pharmacologic perturbations, which can be used to detect novel relationships between genes and small molecules in specific disease by matching the pattern of gene expression changes induced by these agents (114). However, using a genome-scale gene expression array to profile many samples is economically infeasible. Peck and colleagues developed L1000 assay, which is a Luminex bead based gene expression profiling of 978 landmark genes whose collective gene expression signature predicts the expression of all other genes in the genome with high accuracy (115,116). Investigating the gene expression change associated with specific genetic alterations in cancer may allow understanding the functional impact of those alterations.

### 1.4.6 Emerging methods – genome editing

Recent developments in genome editing technologies such as zinc finger nucleases (ZNFs), transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR) are dramatically enhancing our ability to

interrogate genetic alterations in cancer by facilitating recreation of these alterations with much less efforts and resources (117,118). These technologies harness nucleases that allow customizable, precise recognition of target DNA sequences (117). The cut in DNA sequences made by these nucleases can be repaired with endogenous cellular DNA repair machinery. Random repair can results in loss of function of the gene by introducing frameshift mutations. Providing intended repair templates upon cutting can yield repaired DNA sequences with the desired sequence alterations incorporated (117). CRISPR technology, which uses guide RNA complementary in sequence to the target DNA sequences, has been adapted to generate knockout and knock-in mice to study genes involved in tumor development, metastasis and resistance to drug treatment (119-123). Introduction of multiple guide RNAs and templates allows the generation of mice with compound mutations in a single step, bypassing the labor-intensive crossbreeding required for conventional compound transgenic mouse generation. It is expected that knocking in desired genetic alterations using CRISPR technology will be amenable to high throughput adaptation; novel methods to increase the efficiency of CRISPR mediated knock-in are being developed (124-126). However, presently, this technology requires generating one alteration at a time and screening for the correctly altered clones, making it not scalable.

In this thesis, I address the challenges of investigating non-synonymous point mutations and focal amplifications in cancer by utilizing a systematic *in vivo* gain-of-function ORF screen (Chapter 2), gene expression analysis (Chapter 3), and loss-of-function shRNA screens (Chapter 4).

# Chapter 2

# Pooled *in vivo* screen identifies rare oncogenic alleles

This chapter is adapted from:

**Systematic functional interrogation of rare cancer variants identifies oncogenic alleles**

Cancer Discov. 2016 May 4. pii: CD-16-0160. [Epub ahead of print] PMID: 27147599

Eejung Kim*, Nina Ilic*, Yashaswi Shrestha*, Lihua Zou*, Atanas Kamburov*, Cong Zhu, Xiaoping Yang, Rakela Lubonja, Nancy Tran, Cindy Nguyen, Michael S. Lawrence, Federica Piccioni, Mukta Bagul, John G. Doench, Candace R. Chouinard, Xiaoyun Wu, Larson Hogstrom, Ted Natoli, Pablo Tamayo, Heiko Horn, Steven M. Corsello, Kasper Lage, David E. Root, Aravind Subramanian, Todd R. Golub, Gad Getz, Jesse S. Boehm, William C. Hahn

(*co-first author)

**Contribution:**

E.K., N.I., Y.S, G.G., J.S.B. and W.C.H. conceived the study.

M.S.L., and A.K. led cancer genomics analysis for variant selection.

Y.S., C.Z., X.Y., R.L., N.T., C.N. and D.E.R. generated cDNA reagents.

E.K., N.I., Y.S., C.R.C. and J.G.D. performed tumorigenesis experiments

F.P., M.B., X.W, Y.S. and D E.R. performed gene expression experiments.

E.K., L.Z, L.H., T.N. and S.M.C. performed gene expression analyses.

H.H., K.L., P.T. and S.M.C. advised on analysis strategy.

E.K., N.I., A.K., Y.S. and C.R.C. performed validation experiments.

E.K., J.S.B. and W.C.H. wrote the manuscript.

All authors participated on discussion and commented on the manuscript.

**Specific contribution:**

Atanas Kamburov generated Figure 2-7B.

Nina Ilic generated Figures 2-7 and 2-8, except Figure 2-7B.

Eejung Kim generated all other figures in this chapter.

Iris Fung and Bang Wong helped with illustrations in Figures 2-1 and 2-3.

## 2.1 Introduction

Describing the complete list of genes altered in cancer genomes has been a major goal of cancer research, with an expectation that identifying mutated cancer genes would elucidate the molecular basis of cancer and nominate potential therapeutic targets (3). Advancements in sequencing technologies have facilitated the initial description of mutational landscapes in many types of cancers (31,32). Although these efforts have identified oncogenes and tumor suppressor genes that occur at high frequency, the majority of somatically altered alleles are found at low frequency, making it difficult to differentiate functionally relevant alleles from neutral, passenger mutations (31). Computational approaches to predict the functional consequences of these low incidence point mutants are informative but require experimental and clinical validation (55).

Increasing numbers of cancers are now being sequenced in clinical settings, and in some cases this information is used to direct therapeutic decisions (127-130). Although such efforts will facilitate recruitment to clinical trials of molecularly targeted agents, it is already clear that such efforts identify many somatically altered but unstudied alleles in known oncogenes and tumor suppressor genes as well as in genes not previously implicated in cancer initiation or progression (128,131). At present, such alleles are either classified as variants of unknown significance (VUS) or are not reported (132,133).

Although the in-depth study of single genes will eventually provide functional information for these cancer-associated alleles, it is now possible to systematically study the consequences of expressing mutant alleles at scale. To determine whether the systematic characterization of cancer alleles can provide functional insights, we generated a large number of alleles identified in cancer genome sequencing studies and assessed the consequences of expressing these alleles on tumor formation and gene expression (**Figure 2-1**). In this chapter, I describe the curation of mutated alleles selected for this study, and the results of the pooled *in vivo* screen.

In the Chapter 3, I analyze the gene expression data of these mutated alleles as well as wild type and reference alleles with known biological function. These two methods represent a scalable approach to characterize and assign function to a large number of alleles identified by cancer genome sequencing efforts.

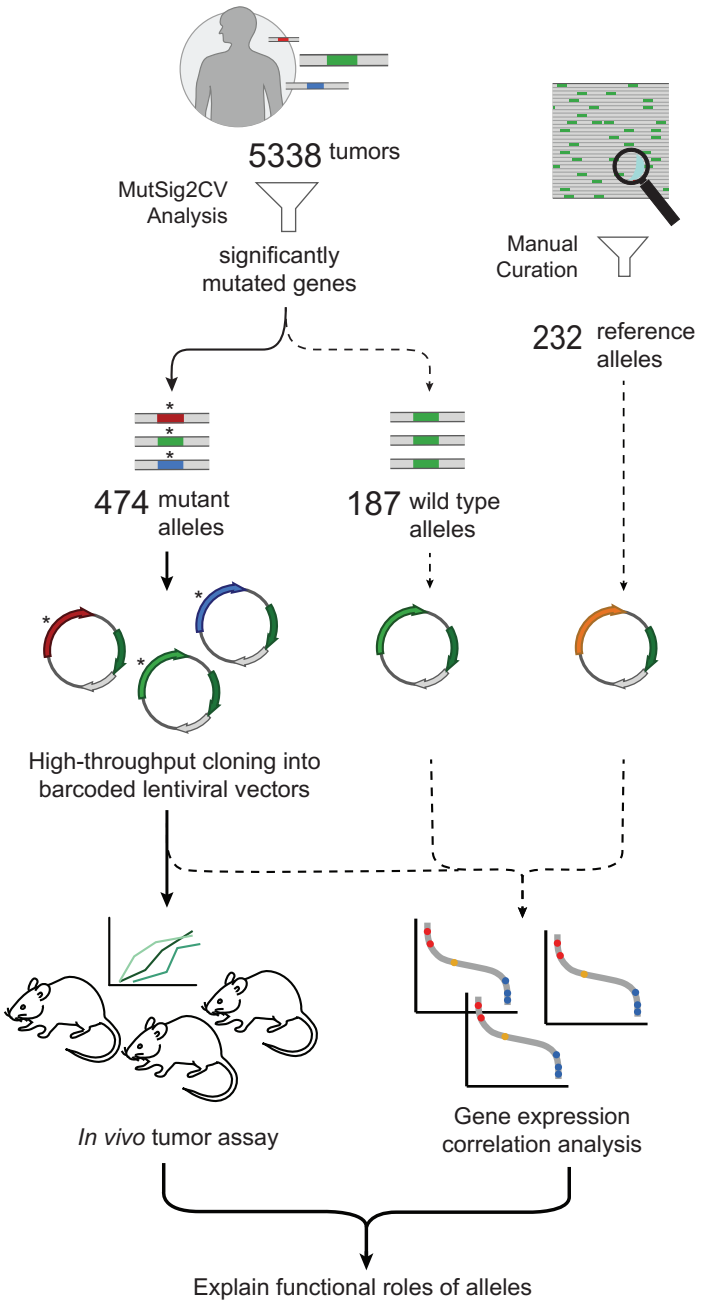



**Figure 2-1. Project pipeline.**

## 2.2 Results

### 2.2.1 Creation of a Pan-Cancer candidate cancer allele panel

To create a panel of cancer alleles, we first identified candidate cancer genes by running MutSig2CV (134,135) on a collection of 5,338 tumors representing 27 cancers that had been subjected to whole exome or whole genome sequencing. Specifically, we prioritized genes by their p-value calculated from their individualized background mutation rate, which was determined by considering covariates such as gene expression level and DNA replication timing (134). These analyses identified 381 genes, 220 for which (58%) templates were present in the hORFeome 8.1 collection of cDNA clones (106) (Supplementary Table S1). We selected 696 mutant alleles for reagent generation by considering local mutational density and evolutionary conservation (described in Materials and Methods). Of the 220 alleles for which we had templates, we generated 187 wild type alleles and 474 of the 696 nominated mutated alleles (68%, 178 genes). In addition, we constructed and included a set of 232 ORFs with known functions as well as 24 control ORFs. These alleles were introduced into uniquely barcoded lentiviral vectors. In total, this collection included 1163 ORFs (Materials and Methods; Supplementary Table S2).

The majority of the 474 mutant alleles were infrequently mutated in human cancers. Specifically, 350 (73.8%) of the mutant alleles were found only once, and 12.0%, and 4.9% of the alleles were found twice and three times, respectively (**Figure 2-2A**). We noted that as the frequency of an allele increased, that allele was more likely to be found in multiple lineages (**Figure 2-2B**). These observations suggest that testing these alleles in a single cell context may provide generalizable information.

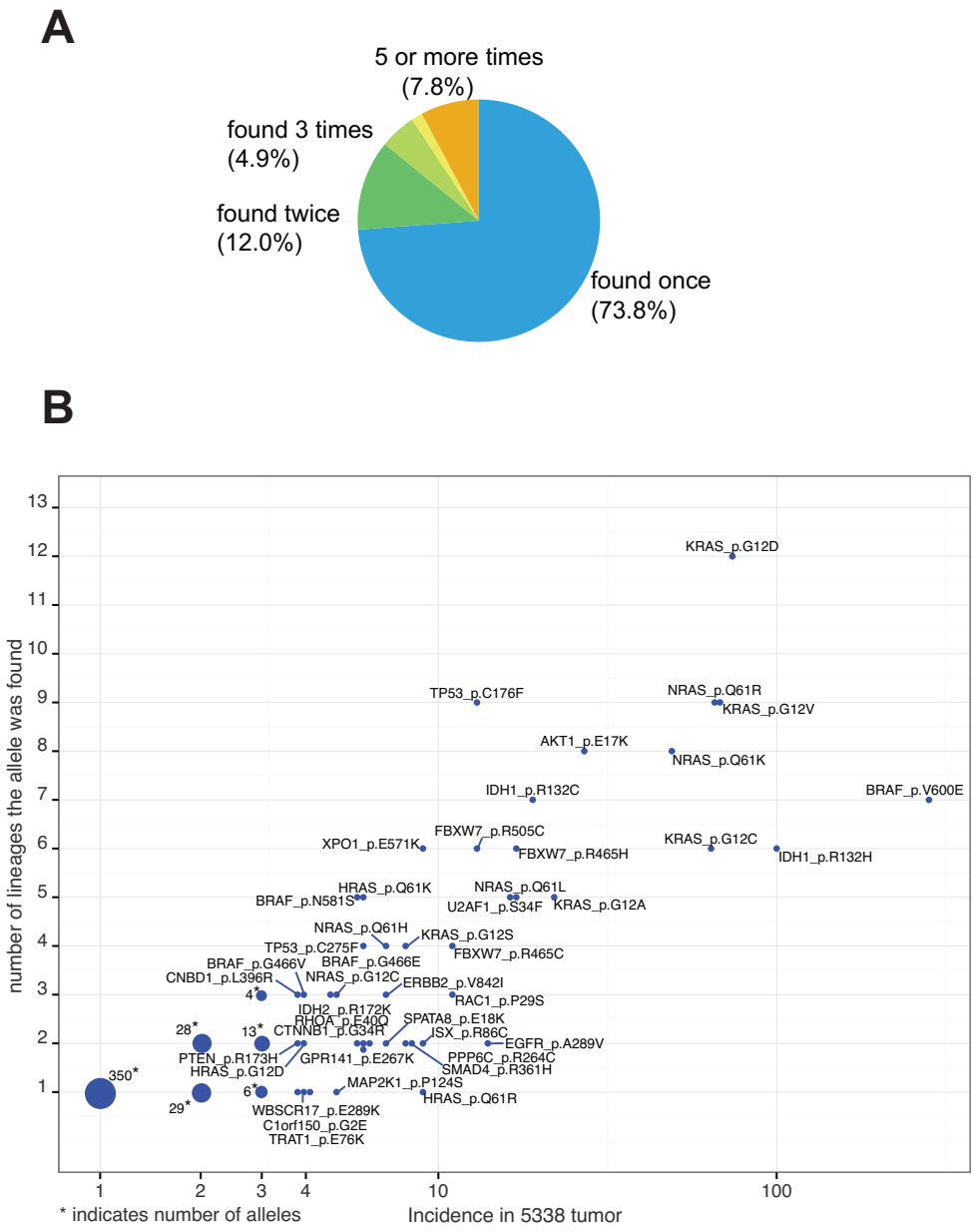



**Figure 2-2. Summary of alleles included in this study**.

**(A)** Distribution of incidence of the alleles included in the project. 73.8% of the 474 alleles included in this study were found to be mutated only once.

**(B)** Alleles mutated frequently were also found to be mutated in larger number of lineages. The size of dots corresponds to the number of overlapping dots.

### 2.2.2 High-throughput identification of transforming alleles in vivo

The assessment of tumor formation potential in mice is a widely used method to assess transforming function of specific alleles. We created a high-throughput platform to determine whether specific cancer-associated alleles induce tumor formation. We chose genetically defined, immortalized human embryonic kidney cell line, HA1E (136), and HA1E cells expressing an activated $MEK1^{DD}$ allele (HA1E-M) as model systems. HA1E-M cells are primed for cell transformation and have been previously used for somatic genetic screens (21,137). We expressed each of the 474 alleles in HA1E-M cells and then used an *in vivo* pooled strategy to assess the tumorigenic potential of each allele (**Figure 2-3**).

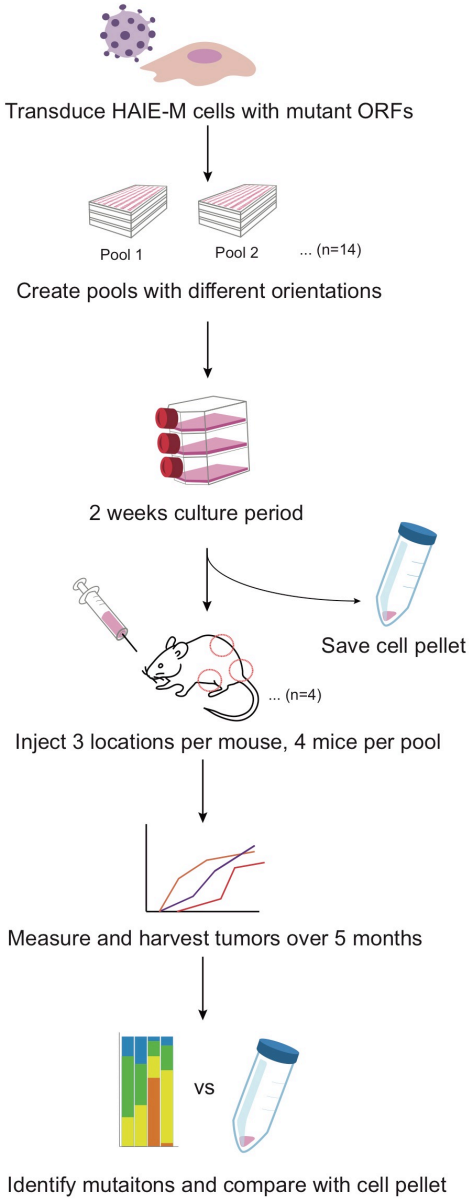



**Figure 2-3. Pooled *in vivo* screen design**.

HA1E-M cells were transduced with lentivirus harboring 474 mutant constructs in arrayed fashion. These 474 different cell lines were pooled in 14 pools and each pool of cells was injected into three sites on four immunocompromised mice. The injection sites were observed for five months for tumor development. Allele compositions of the cell pellets and each tumor were compared for enrichment and penetrance calculations.

Based on optimization experiments, we placed all 474 alleles into seven different pools (Pools 1-7) and segregated known oncogenic alleles into Pool 1, to reduce the possibility that known transforming alleles would dominate tumor formation and mask weaker oncogenic alleles. Pool 8 is a biological replicate of Pool 1. We scrambled alleles in Pools 2-7 into Pools 9-14 to create an additional set of pools, to give each allele two different sets of pool neighbors to increase sensitivity. The pool composition is described in Supplementary Table S3. We transduced each of the alleles into HA1E-M cells in an arrayed format, then pooled and expanded cells for tumorigenicity studies (**Figure 2-3**; Materials and Methods). Barcode sequencing of ORFs confirmed that nearly all of the alleles were represented upon implantation, although we noted that the representation of the alleles was not equal, likely due to the differences in viral titer because of differences in the length of each ORF and nucleotide composition (**Figure 2-4**).

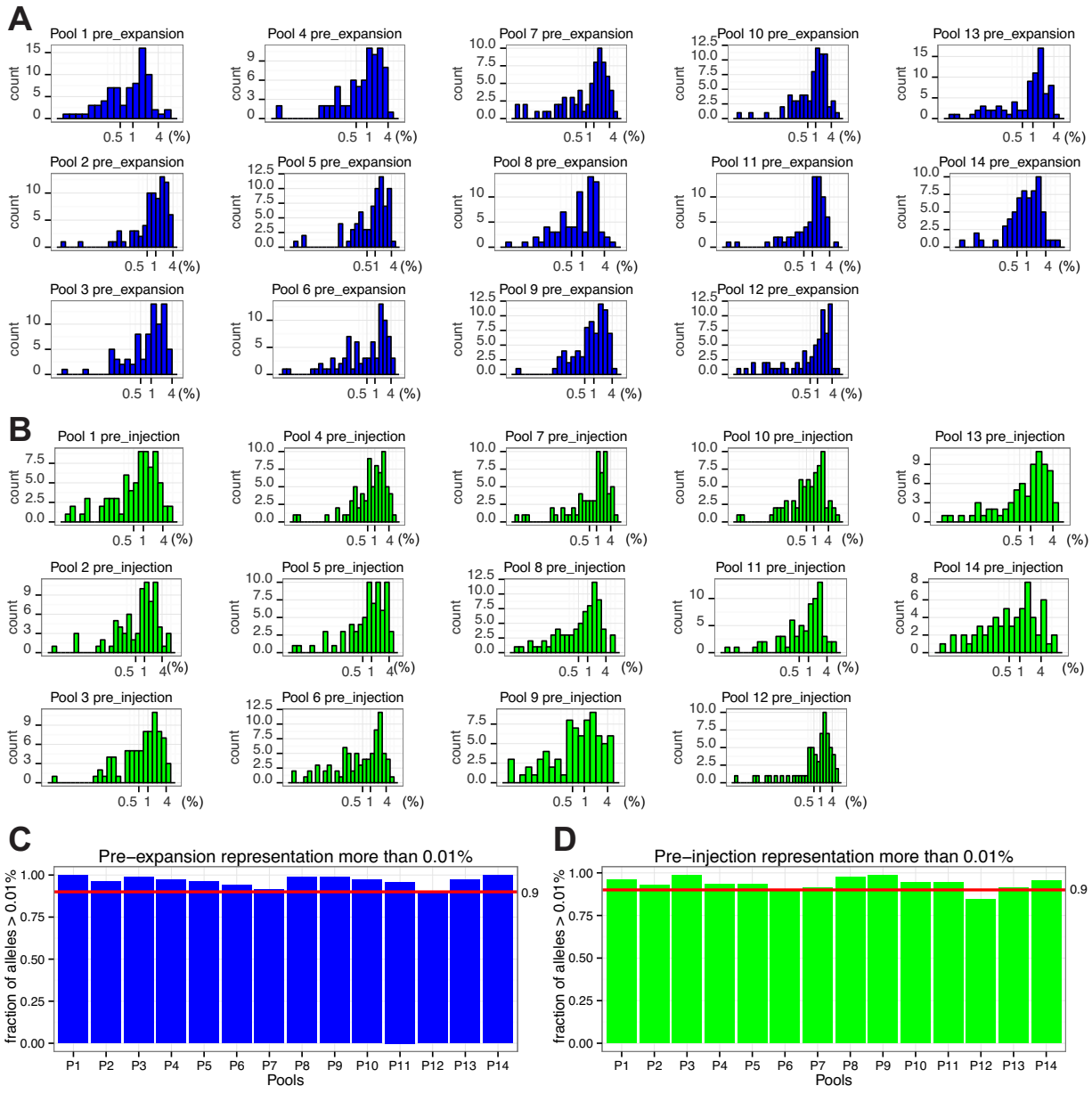



**Figure 2-4. Distribution of barcode read representation in pre-expansion and pre-injection samples.**

(**A**) Allele representation immediately after pooling cells (called “pre-expansion”) according to the pool composition (Supplementary Table S3). Each pool contains ~75 alleles. The majority of alleles were represented at 0.5-4%. The data for this histogram is available in Supplementary Table S4-1.

**Figure 2-4 (Continued).**

(**B**) Allele representation after 15-day culture, immediately before the injection into nude mice (called “pre-injection”). The majority of alleles were represented at 0.5-4%. The data for this histogram is available in Supplementary Table S4-2.

(**C**) Percentage of alleles in each pool that was represented at more than 0.01% in pre-expansion cell pellet.

(**D**) Percentage of alleles in each pool that was represented at more than 0.01% in pre-injection cell pellet.

Pools consisting of known cancer alleles (Pools 1 and 8), formed tumors within 1-2 weeks (**Figure 2-5**), and all eight mice in these pools were sacrificed by week 3. Pools 7 and 14, experimental pools with a total of 110 unique alleles, failed to form any tumors after 18 weeks, confirming previous work showing that the background rate of tumor formation is low in this experimental model (**Figure 2-5**). We harvested 69 tumors from 168 implantation sites and quantified the barcodes associated with each ORF by PCR amplification and sequencing (Materials and Methods; Supplementary Table S4).

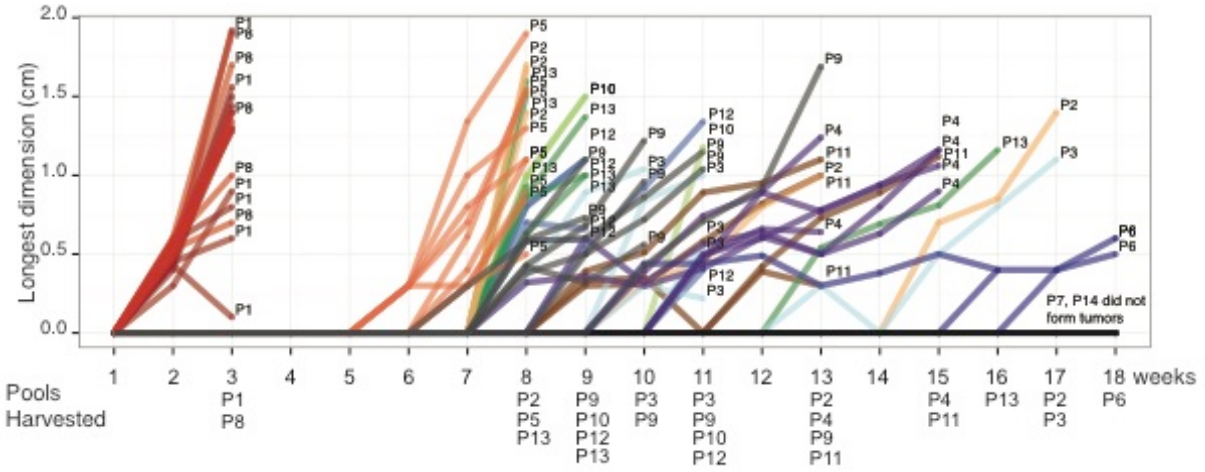



**Figure 2-5. Tumor formation over an 18-week timeframe per pool.**

Positive control pools (P1, P8) formed tumors within three weeks of injection. Tumors from other pools tended to form tumors around the same time. Tumors from some pools (Pool 6) never grew bigger than 0.7cm in the longest dimension.

We observed that tumors derived from pools 1 and 8, which were composed of known oncogenic alleles, repeatedly demonstrated a similar pattern of allele representation, mainly composed of *NRAS* and *KRAS* alleles (**Figure 2-6A**). In contrast, we found that tumors derived from other experimental pools showed a wide diversity of allele representation. Some pools contained a single dominant oncogenic allele while others included several oncogenic alleles (**Figure 2-6B, C, D**). Certain alleles, such as $KRAS^{D33E}$, were found enriched in all tumors in which they were assessed; we labeled these alleles as highly penetrant (**Figure 2-6E**). Other alleles such as $POT1^{G76V}$ were less penetrant but they were highly enriched in a few tumors (**Figure 2-6C, E**). We noted that the $KRAS^{A59G}$, $AKT1^{L52R}$, $AKT1^{Q79K}$, $NFE2L2^{G31R}$, $NFE2L2^{WT}$, $PIK3CB^{E497D}$, $FAM200A^{S481N}$ alleles were found at more than 1% in at least two tumors in the pooled screen (**Figure 2-6E**; **Figure 2-7**).

The pooled nature of the screen forces competition among alleles in the same pool. For example, Pool 1, only eight alleles out of 77 were represented at 1% or higher in tumors and

when lower threshold of 0.01% was applied, 24 alleles met the cutoff (Supplementary Table S4-3). Known oncogenic alleles such as *AKT1*$^{E17K}$ failed to score due to competition, even though this allele is known to transform in this cellular context (137). Nevertheless, these observations allowed us to identify a subset of somatically altered alleles that induce tumor formation in this context.

**Figure 2-6. Tumor composition of *in vivo* pooled screen and summary**.

(**A**) Pool 1, a positive control pool, showed consistent tumor composition across tumors. Each tumor is represented as a bar. The compositions of tumors were shown as different colors.

(**B**) *KRAS*$^{D33E}$ induced tumor formation in pool 5.

(**C**) *NFE2L2*$^{G31R}$ and *POT1*$^{G76V}$ induced tumor formation in pool 4.

(**D**) *NFE2L2*$^{G31R}$ and *PIK3CB*$^{E497D}$ induced tumor formation in pool 9.

(**E**) Summary of the *in vivo* pooled screen. X-axis shows penetrance, which was calculated to be (times each allele was more than 0.01% of tumor reads) / (number of sites the allele was implanted). Since mice must be sacrificed when the largest tumor reaches a threshold, not all sites were observed for the full length of time. Y-axis shows maximum enrichment, which was calculated to be (maximum percentage of allele in any tumor) – (percentage of the allele in pre-injection cell pellet). Positive controls (colored in grey) had penetrance of around 80%, and low maximum enrichment due to competition against each other.

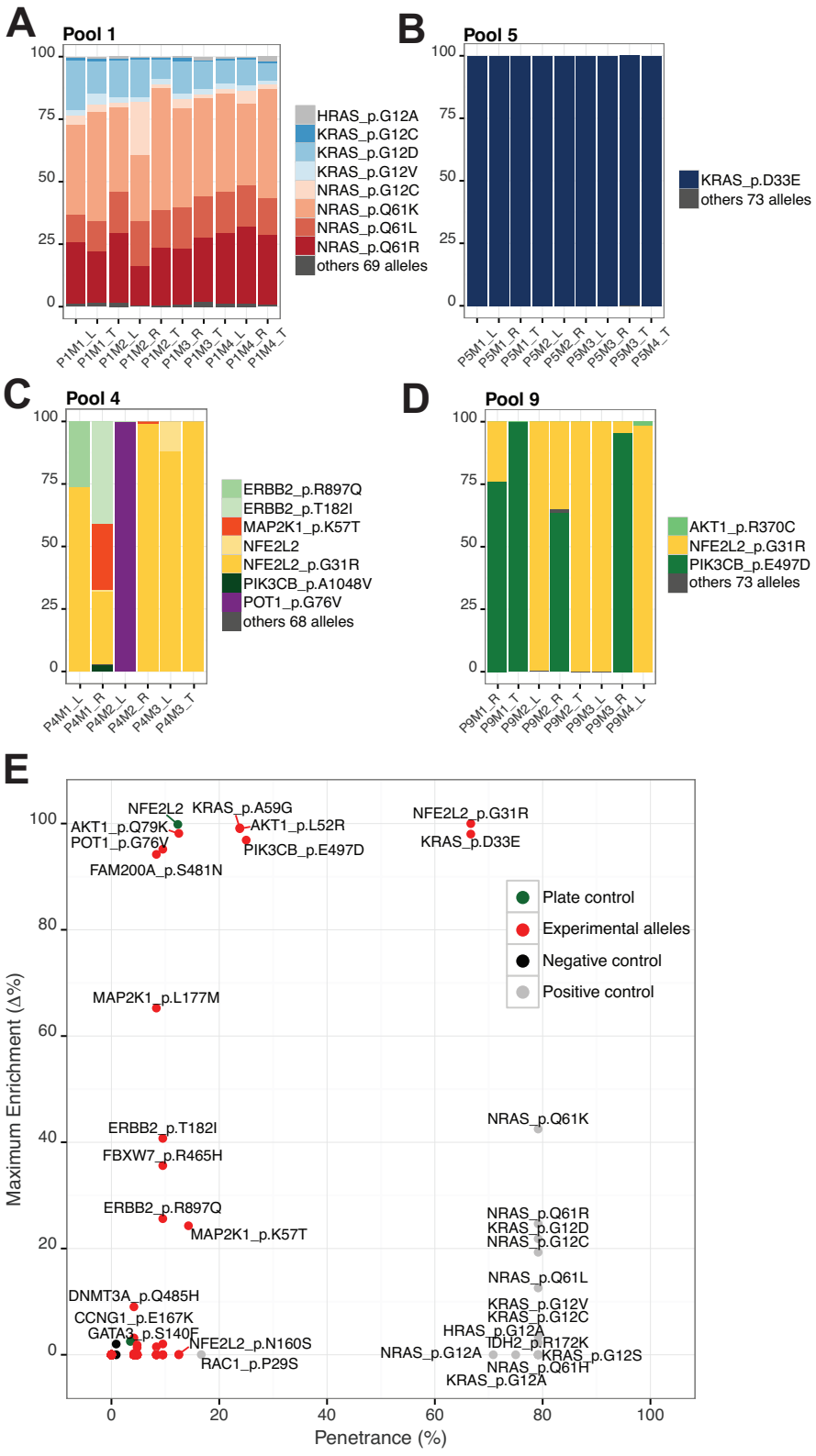



**Figure 2-6. (Continued).**

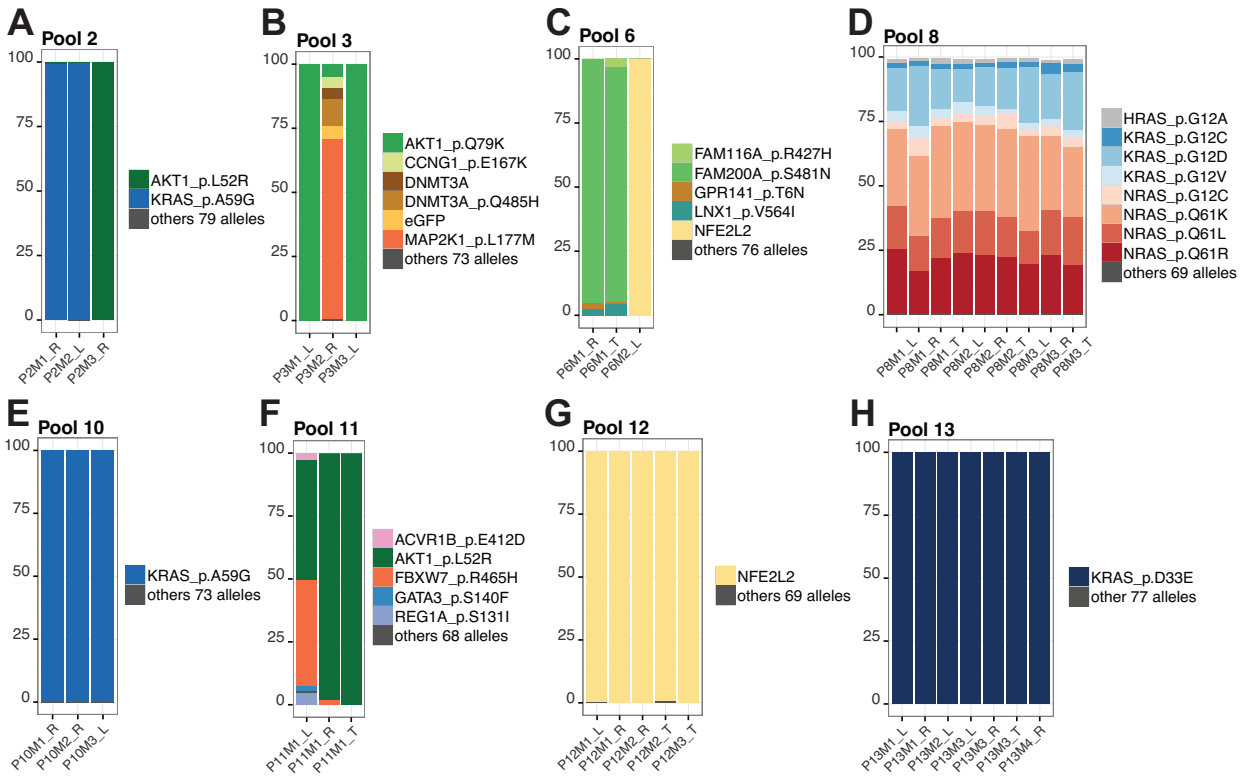



**Figure 2-7. Tumor composition of *in vivo* pooled screen**

(**A**) Tumor composition of pool 2. *AKT1*$^{L52R}$ and *KRAS*$^{A59G}$ scored.

(**B**) Tumor composition of pool 3. *AKT1*$^{Q79K}$ and *MAP2K1*$^{L177M}$ scored.

(**C**) Tumor composition of pool 6. *FAM200A*$^{S481N}$ and *NFE2L2*$^{WT}$ scored.

(**D**) Tumor composition of pool 8. Tumor composition was analogous to that of pool 1.

(**E**) Tumor composition of pool 10. *KRAS*$^{A59G}$ scored.

(**F**) Tumor composition of pool 11. *AKT1*$^{L52R}$ and *FBXW7*$^{R465H}$ scored.

(**G**) Tumor composition of pool 12. *NFE2L2*$^{WT}$ scored.

(**H**) Tumor composition of pool 13. *KRAS*$^{D33E}$ scored.

### 2.2.3 Validation of rare oncogenic alleles

To validate the tumor formation of rare alleles, we performed individual tumorigenicity experiments with the candidate oncogenic alleles and their allelic series (**Figure 2-8**, **Figure 2-9**). We defined tumorigenic allele in as an allele that formed any tumor larger than 500 $mm^3$ by 130 days. We validated that *AKT1*$^{L52R}$, *NFE2L2*$^{G31R}$, *POT1*$^{G76V}$, *KRAS*$^{D33E}$, and *KRAS*$^{A59G}$ were tumorigenic. In addition, some alleles that did not score in pooled screen formed tumors in individual experiment including *KRAS*$^{E62K}$, *PIK3CB*$^{A1048V}$, *NFE2L2*$^{G31A}$, *NFE2L2*$^{G31V}$, *NFE2L2*$^{N160S}$, *AKT1*$^{E267G}$ and *AKT1*$^{R370C}$ (**Figure 2-8A, E, H, J**).

We found that the *KRAS*$^{D33E}$ and *KRAS*$^{A59G}$ alleles were potently tumorigenic, while the *KRAS*$^{E62K}$ allele induced tumor formation at much longer latencies (**Figure 2-8A**). When we mapped the *KRAS*$^{D33E}$, *KRAS*$^{E62K}$, and *KRAS*$^{A59G}$ on the KRAS structure (138) we found that these mutations occur in close proximity with known transforming alleles (**Figure 2-8B**). Cells expressing *KRAS*$^{D33E}$ and *KRAS*$^{A59G}$ showed increased activation of the MAP kinase and PI3K pathways as assessed by phosphorylation of specific effectors and a RAF binding domain pull down assay (**Figure 2-8C, D**). These observations suggest that these rare *KRAS* alleles are indeed oncogenic.

When we examined the *NFE2L2* allelic series, we found that the G31R, G31V, G31A, and T80K alleles robustly formed tumors (**Figure 2-8E**), while the N160S allele formed small tumors at a much later time point. We note that expression of wild type *NFE2L2* induced the formation of a single tumor formation at long latency. Tumor formation by *NFE2L2* wild type overexpression was also observed in the pooled screen (**Figure 2-7G**). In consonance with these observations, we found that tumorigenic *NFE2L2* mutants were expressed a higher levels, likely due to impaired degradation mediated by endogenous KEAP1 (**Figure 2-7F, G**).

In individual tumor assays, *PIK3CB*$^{E497D}$ showed delayed tumor formation, similar to what we observed when we expressed the wild type *PIK3CB* (**Figure 2-7H**), implying E497D is

a passenger mutation. Wild type *PIK3CB* was previously shown to induce foci in a foci formation assay (139). *PIK3CB*$^{A1048V}$, on the other hand, induced tumors in the majority of replicates with shorter latency, demonstrating that *PIK3CB*$^{A1048V}$ is a transforming gain-of-function mutant. In the *POT1* allelic series, we noted that only *POT1*$^{G76V}$ formed tumors in individual tumor experiments after long latency. *POT1* was recently shown to be mutated in familial melanoma (140,141), chronic lymphocytic leukemia (142), familial glioma (143), and cardiac angiosarcoma (144). In particular, Y89C, Q94E, R273L, Y223C, and S270N alleles were previously shown to be loss-of-function, resulting in elongated telomeres and increased genomic instability (140,141). These observations suggest that *POT1*$^{G76V}$ may also contribute to cell transformation through a similar mechanism.

Although some of the alleles that we found induced tumor formation were recurrently observed in particular human cancer types, we noted that many of the alleles that we found were able to induce tumor formation, including *KRAS*$^{D33E}$, *KRAS*$^{E62K}$, *NFE2L2*$^{G31R}$, *NFE2L2*$^{G31V}$, *NFE2L2*$^{N160S}$, *POT1*$^{G76V}$ and *PIK3CB*$^{A1048V}$, were found to be mutated only once in our set of 5,338 tumors. These observations demonstrate that rare alleles may be functionally important in tumorigenesis.

**Figure 2-8. Validation of rare oncogenic alleles in *KRAS*, *NFE2L2*, *PIK3CB* and *POT1*.**

(**A**) Individual tumor validation of *KRAS* alleles. The *KRAS*$^{D33E}$ and *KRAS*$^{A59G}$ alleles formed tumors robustly. E62K did not form tumors in the pooled assay but formed tumor in individual assays, at a later time point.

(**B**) The structure of KRAS (PDB: 4EPV) shows that all four of the mutants are spatially close. Mutated residues are shown in red, GDP bound to the substrate pocket is shown in blue.

(**C**) Immunoblot of KRAS alleles (including other positive control alleles) shows increased phospho-ERK and phospho-AKT1 levels in *KRAS*$^{D33E}$, and *KRAS*$^{A59G}$ overexpressed cells.

(**D**) RAF binding domain pull down assay shows increased GTP bound KRAS in D33E and A59G mutants.

(**E**) Individual tumor validation of *NFE2L2* alleles. In the pooled assay, only G31R scored in multiple tumors. In the individual assay, G31V, G31A, T80K formed tumors as well. N160S formed tumors at a later time point. *NFE2L2* wild type formed one small tumor by the end of the experiment.

(**F**) Quantitative PCR of NFE2L2 mRNA expression shows all alleles were expressed.

(**G**) Immunoblot of *NFE2L2* alleles show increased NFE2L2 protein level in G31A, G31R, G31V and T80K overexpressed cells. There was no change in phospho-ERK or phospho-AKT1 levels.

(**H**) Individual tumor validation of *PIK3CB* alleles. E497D and the wild type formed tumors after long latency. *PIK3CB*$^{A1048V}$ formed tumors with shorter latency at the majority of injection sites.

(**I**) Individual tumor validation of *POT1* alleles. The G76V allele formed tumor at a later time point. One of the *POT1*$^{G76V}$ mice died of unknown cause.

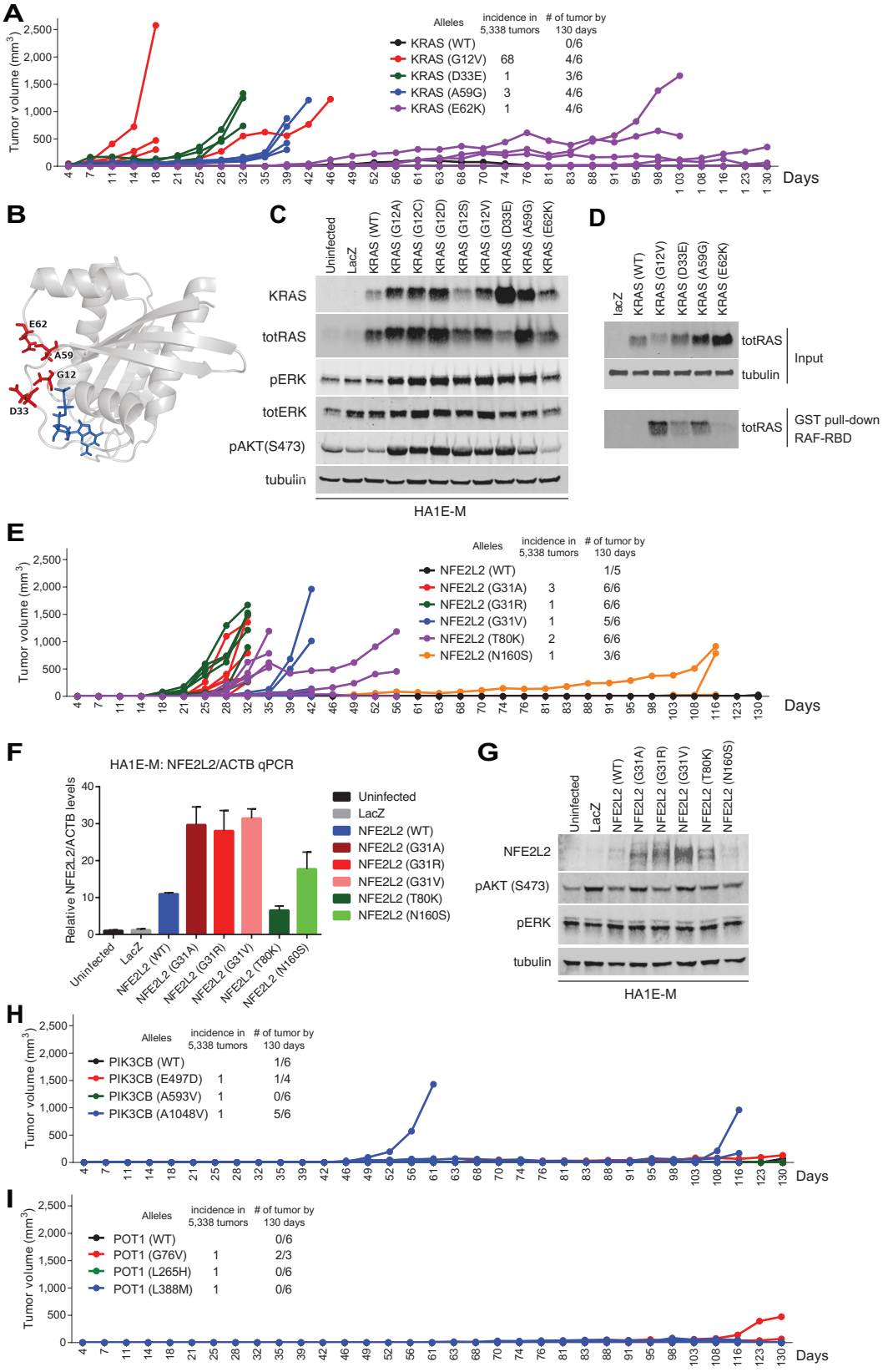



**Figure 2-8. (Continued).**

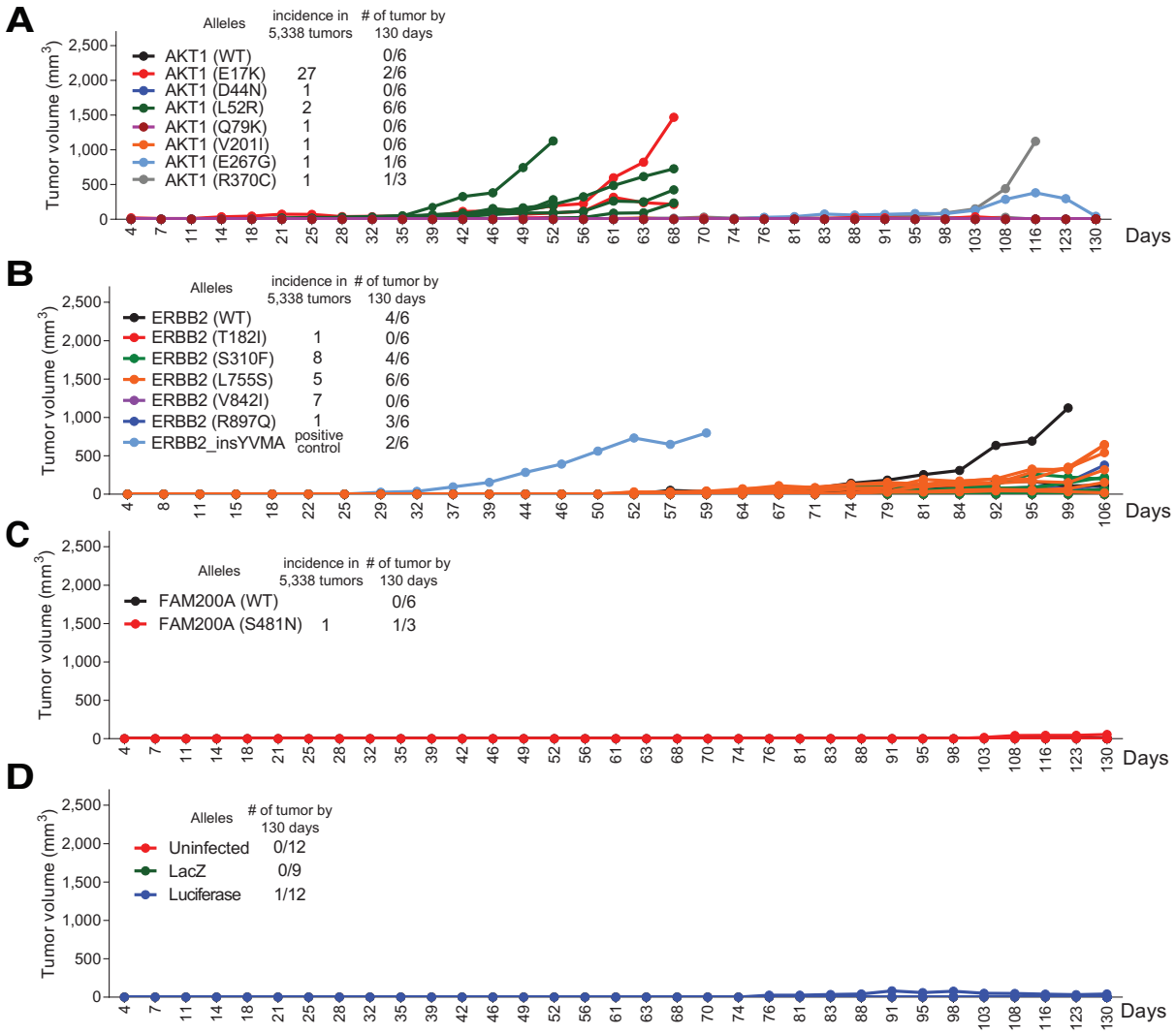



**Figure 2-9. Validation of rare oncogenic alleles in *AKT1*, *ERBB2*.**

(**A**) Individual tumor validation of *AKT1* alleles. E17K, L52R, E267G, and R370C formed tumors. Q79K did not form tumor. One mouse of *AKT1*$^{R370C}$ died of unknown reason.

(**B**) Individual tumor validation of *ERBB2* alleles. InsYVMA mutant was included as a positive control, which was described previously (145). Tumor forming alleles formed tumors in a similar timeframe to that of the wild type.

(**C**) Individual tumor validation of *FAM200A* alleles. *FAM200A*$^{S481N}$ formed one small tumor at later time point. One mouse of *FAM200A*$^{S481N}$ died of unknown reason.

(**D**) Negative controls in individual tumor validation. Four mice were used in each of uninfected,

**Figure 2-9. (Continued).**

LacZ-transduced, and Luciferase-transduced groups. One small tumor formed in Luciferase-transduced groups and regressed spontaneously. One mouse in LacZ-transduced group died of unknown reason.

## 2.3 Discussion

Cancer genome sequencing projects have identified thousands of variants of unknown significance, and this number will likely increase rapidly as more tumors are sequenced. Here we report a pilot study to facilitate the functional characterization of these alleles by creating a large number of cancer-associated variants and testing them in an *in vivo* tumorigenesis assay. We identified a subset of these variants that exhibit tumorigenic phenotypes. This study provides proof of principle evidence that large-scale mutant characterization is both tractable and provides new information about the functional relevance of many alleles.

We recognize that these studies are not exhaustive. For example, we performed all experiments using immortalized kidney epithelial cells, thus limiting those genes that are potentially transforming in a specific tissue context. In addition, the tumorigenesis assay we used here does not assess all tumor-essential functions and this experimental design does not permit the discovery of loss-of-function tumor suppressor alleles. For example, alleles involved in metastasis, angiogenesis, immune response, and splicing changes may not score in this assay. Weaker transforming alleles may be masked by stronger oncogenic alleles in the pooled format used in these experiments and it is possible that there are both productive and inhibitory interactions between cells harboring different alleles. Furthermore, alleles that affect pathways that were already perturbed in our engineered system, which include inhibition of *TP53* and *RB* as well as *hTERT* and *MEK*$^{DD}$ overexpression, are not likely to be discovered in this context.

Also, in cases where presumable mechanisms involve stochastic accumulation of mutations over long time periods, as in the case of genes involved in genomic instability such as *POT1*, these genes may not reliably score in this context. However, considering the very low background tumor formation rate in this assay, even a single instance of tumor formation lends support for future studies. As such, this approach provides a powerful paradigm to discover functionally relevant rare alleles that may otherwise not be considered for functional studies due to their rarity. Further studies such as those described herein using similar approaches in other genetic and lineage contexts will facilitate the comprehensive discovery of transforming alleles.

Using the *in vivo* tumorigenesis assay, we identified rare mutants with transforming function, such as $KRAS^{D33E}$. As this variant was identified only once in the cohort of 5,338 tumors, a large number of tumors would need to be sequenced before the frequency of this allele reached statistical significance. As KRAS mutational status is already used in directing therapeutic decisions (146), this observation demonstrates the importance of studying rare alleles for accurate patient stratification. $PIK3CB^{A1048V}$ and $POT1^{G76V}$ were also rare alleles that were found only once in our cohort. *PIK3CB* was recently shown to be mutated in prostate cancer (147), and computational analysis using network mutation burden nominated *PIK3CB* to be a significantly mutated gene (Horn et al, submitted). Although further studies are required to elucidate the mechanisms by which $PIK3CB^{A1048V}$ and $POT1^{G76V}$ contribute to malignant transformation, this study provides evidence that these alleles are indeed transforming alleles.

In this study, we focused on alleles that have been identified in cancer genome sequencing efforts. An alternative approach would be to create a set of alleles where each amino acid is substituted to prospectively identify alleles that alter wild type gene function and to interrogate the relationship among evolutionary conservation, gene function and prevalence of mutations in tumors. Although this type of study is not yet feasible at the scale presented here, our studies suggest that expanding the number of alleles in genes will provide useful information. We acknowledge that arbitrarily limiting the number of alleles per gene, especially

in known cancer genes, excluded some well-studied alleles. Including additional criteria, such as 3D spatial clustering (148), may increase the sensitivity of discovering functional alleles. Expanding the number of alleles in genes, especially those already used in clinical decision-making, is also desirable. Furthermore, high throughput adaptation of other functional assays, such as experiments that quantify morphologic changes as well as proteomic and epigenetic differences will expand our knowledge of the functional consequences of mutant alleles.

## 2.4 Materials and Methods

**Mutated gene curation**

271 mutated genes were called from the analysis of 5,338 tumor normal pairs by running MutSig2CV and setting the q-value cutoff at 0.1. The algorithm was described previously (135). 13 genes were manually added (*PIK3C2G*, *PIK3R2*, *PIK3CG*, *PIK3C2B*, *PIK3CB*, *PIK3C2A*, *PIK3R4*, *BCL2*, *BCL3*, *BCL6*, *BCL9*, *BCOR*, *ISX*). 49 likely false positive genes (genes with high background mutation rate) and 48 randomly chosen, likely neutral genes were added. Total of 381 genes were selected for the project. 220 of these genes had matching template in the hORFeome 8.1 collection and these were used for subsequent steps (Supplementary Table S1).

**Selection of alleles from significantly mutated genes**

For each missense mutation, "priority" was calculated, which was defined as "density" (local concentration of mutations) multiplied by conservation.

priority = mutation density * conservation

Mutation density was calculated by counting the number of mutations in 20bp window, with the allele of interest at the center of the window. Conservation was calculated by using phyloP

(149), which scores evolutionary conservation from an alignment of 46 vertebrates. Conservation values were scaled linearly to range from 0 to 100.

We chose an allele by taking the highest-priority mutated allele. The same procedure was repeated until we selected as many alleles as desired. The number of alleles selected for each gene was decided by the number of times the gene was mutated in patients.

1. If a gene was mutated in 120 patients or more, then 8 alleles were chosen.
2. If a gene was mutated in 100 patients or more, then 7 alleles were chosen.
3. If a gene was mutated in 80 patients or more, then 6 alleles were chosen.
4. If a gene was mutated in 70 patients or more, then 5 alleles were chosen.
5. If a gene was mutated in 60 patients or more, then 4 alleles were chosen.
6. If a gene was mutated in 50 patients or more, then 3 alleles were chosen.
7. If a gene was mutated in 30 patients or more, then 2 alleles were chosen.
8. Otherwise, one allele per gene was chosen.

For *HRAS*, *SPOP*, *MAP2K1*, *B2M*, *AKT1*, *RHOA*, *IDH1*, and *IDH2*, 8 alleles were chosen.

For genes with one or two alleles selected, we considered all the mutations as 'experimental' alleles. For genes with three or more alleles selected, we selected one allele that we predicted to be less likely to be functional as a 'control' allele. The other alleles were considered 'experimental' alleles. The ‘control’ allele was chosen as an internal control that is less likely than the ‘experimental’ alleles to be functional. The ‘control’ alleles were chosen by the following criteria.

1. Remove any positions that were chosen above.
2. Remove any mutations with conservation above a threshold of 60.
3. For the remaining mutations, define controlpriority = (100 - conservation) / (# of times that exact mutation occurs)^2.
4. Add a bonus for mutations that are close to the first or second mutations chosen above. If distance between first or second experimental allele and the control allele was less than one

fifth of the total protein length, bonus of 20 was given. If distance between first or second experimental allele and the control allele was less than one third of the total protein length, bonus of 10 was given.

5. Choose the mutated allele with the highest controlpriority + bonus.

All selected alleles are shown in Supplemental Table S1.

**Barcoded mutant allele generation in lentiviral vectors**

We used a previously published method to perform high-throughput mutagenesis (150). Briefly, each ORF was PCR amplified by using primers that contain mutated sequence incorporated. These fragments were transferred to pDONR223 vector (Invitrogen) through BP cloning (Invitrogen) and the constructs were transformed into competent cells. The discontinuity at the mutation introduction site was repaired by endogenous bacterial repair mechanism. The mutated ORF was transferred to the barcoded destination vector by LR reaction (Invitrogen).

**Lentivirus generation**

Virus were prepared according to the RNAi Consortium (TRC) virus protocol (http://www.broadinstitute.org/rnai/public/resources/protocols).

**Cell lines**

HA1E-M and HA1E cells were previously described (137). Both cell lines were cultured in MEM-alpha (Invitrogen) with 10% FBS (Sigma-Aldrich) and 1% penicillin/streptomycin (Gibco) supplementation. Both cell lines tested negative for mycoplasma.

**Multiplexed *in vivo* screening**

To determine whether the number of cells transduced with a certain allele in a pool of about 80 alleles is sufficient to form tumors, we performed serial dilution and subcutaneous

injection with activating *KRAS* allele, G12V, and found that 1/96 dilution (about 20,000 cells) was still sufficient in forming tumors in all injection sites. For the screen, 2,500 HA1E-M cells were plated in 100ul of media per well in a 96-well plate on day 1. On day 2, polybrene was added to a final concentration of 4ug/ml and 12ul of arrayed viral supernatant was added to the target cell plates. Plates were spun at 2,250 rpm for 30 min at room temperature. After 4 hours, media was changed. After 18 hours, puromycin was added to a final concentration of 2ug/ml. After 48 hours of puromycin selection, cells were trypsinized and pooled. 96 wells were combined into one pool per pool composition (Supplementary Table S3). Cell pellets were taken immediately after pooling (called “pre-expansion”), and also on day 15 to use as a reference points for future analysis. Transduced HA1E-M cells were propagated for 15 days to obtain at least 60 million cells per pool. More than 90% of the ORFs in each pool were represented at 0.01% of the injected cell population (Supplementary Fig S1C, D). We note that alleles with even lower representation, such as *NFE2L2*$^{G31R}$ at 0.0089% in pre-injection cell pellet of Pool4, were sufficient in forming multiple tumors.

On day 15, cells were trypsinized, washed, and counted (called “pre-injection”). Five million cells were prepped in 200ul of PBS per injection site, except pools 2 and 11, for which 4 million cells were prepped per site. Three sites—inter-scapular area, right and left flanks—were injected in each mouse and four mice were injected per pool (12 sites per pool). Mice were monitored for tumor formation and the longest dimension of each tumor was measured. Tumors were harvested when they reached around 2cm. The tumor tissue was finely minced and subjected to genomic DNA extraction with Qiagen DNeasy blood and tissue kit.

1ug of genomic DNA was subjected to PCR amplification for barcode de-multiplexing by sequencing. To amplify the barcodes with Illumina sequencing primer integrated, following primers were used (different sequence components are demarcated by “<>”):

P5 ORF primer:

<P5 flow cell attachment sequence><Illumina sequencing primer><Vector primer binding>

<AATGATACGGCGACCACCGAGATCT><ACACTCTTTCCCTACACGACGCTCTTCCGATCT[s]><TCTT
GTGGAAAGGACGA>

P7 ORF primer:

<P7 flow cell attachment sequence><Barcode><Illumina sequencing primer><Vector primer binding>

<CAAGCAGAAGACGGCATACGAGAT><NNNNNNNN><GTGACTGGAGTTCAGACGTGTGCTCTTCCG
ATCT><TAAAGCAGCGTATCCACATAGCGT>

Upon amplification, the PCR products were purified with AMPure beads and subjected to Illumina sequencing. On average, 1.6 million reads were obtained per tumor.

**In vivo screening analysis**

The barcode reads were de-multiplexed by custom scripts. Less than 1% of contaminating reads (barcode reads that do not belong to the specific pool) were found and removed. The rest of the reads were normalized by dividing the number of reads by the total number of reads from the tumor. Penetrance was calculated by (number of times in which specific allele was represented at more than 0.01%) / (number of times that allele was injected). Since the mouse needs to be sacrificed when the biggest tumor reaches certain diameter per protocol, not all three sites per mouse were observed for full 18 weeks. Maximum enrichment was calculated by (maximum percentage of tumor reads each allele accounted for) – (percentage of that allele in pre-injection cell pellet).

**Stable cell line generation for validation**

For individual validation experiments**,** the same vector used for the pooled screen was used to generate lentiviruses. 80,000 293T cells were plated in one well of 6-well plates. Delta8.9 (900ng), vsv-g (100ng), the ORF vectors (1ug) were transfected in 3ul of TransIT-LT1 Transfection Reagent (Mirus Bio). The viral supernatant was collected after 48 hrs and was

frozen at -80C until use. HA1E-M cells were plated in 6-well plate at 100,000 cells per well. HA1E-M cells were transduced with 300ul of viral supernatant in 8ul/ml polybrene and were spin-infected at 2250rpm for 30minutes. The next day, the media was changed to selection media (puromycin 2ug/ml). After 48 hrs of selection, cells were cultured in puromycin free MEM-alpha complete media (Invitrogen).

**Screen validation**

Six-week old male homozygous NCR-Nu mice (Taconic) were used for xenograft experiments. HA1E-M cell lines stably expressing individual candidate alleles were injected at two million cells per site, except for NFE2L2 alleles, which were injected at one million cells per site. Each stable cell line was injected at three sites per animal, and into two animals, with the total of six sites per cell line. Tumor formation was monitored using calipers twice weekly for 130 days (or 106 days for ERBB2 alleles). Tumor volume was calculated as ((tumor length)*(tumor width)^2))/2.

**KRAS structure analysis**

*KRAS* mutations of interest were overlaid onto the structure of the protein product (PDB: 4EPV) and visualized the structure using PyMOL (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.).

**Immunoblots**

Protein lysates were resolved on 7.5, 4-12, or 8-16% polyacrylamide SDS gels (Bio-Rad), transferred onto nitrocellulose membranes (Bio-Rad) using standard wet-transfer procedures, and incubated with primary antibodies as indicated. All immunoblot assays were visualized using a LI-COR Odyssey infrared imager. The following antibodies were used:

KRAS (Proteintech Group 12063-1-AP), RAS (CST 3965), RAS (clone 10, EMD Millipore 05-516), pERK (CST 4370), ERK (CST 9102), pAKT (S473, CST 4060), a-tubulin (Sigma Aldrich, clone DM1A, T9026), NRF2 (CST 12721), and NRF2 (R&D Systems AF3925) (CST: Cell Signaling Technologies). Secondary anti-rabbit and anti-mouse IRDye antibodies were from LI-COR Biosciences.

**RAS activation assay**

RAS activation assays were performed according to the manufacturer's protocol (Millipore 17-218). In brief, cells were cultured on 6-well dishes and harvested for lysates. A sample of each lysate was saved for input (total RAS load) and the remaining lysate was rocked with glutathione-sepharose 1:1 RAF-RBD slurry in lysis buffer for 1 hour at $4^0$C. The beads were then washed three times with ice-cold lysis buffer, followed by addition of Laemmli/SDS buffer to elute the bound proteins. The RAS-GTP pull-down samples were loaded and resolved on 12% polyacrylamide SDS gels (Bio-Rad).

**Quantitative real-time PCR (qPCR)**

RNeasy kit (Qiagen) was used to purify total RNA from cells and cDNA was generated using Superscript III Vilo (Life Technologies). Quantitative real-time PCR was performed using SYBR reagents (Life Technologies) on an ABI-7300 instrument following a two-step cycling protocol with the following primers:

NFE2L2_FWD: CACATCCAGTCAGAAACCAGTGG

NFE2L2_REV: GGAATGTCTGCGCCAAAAGCTG

ACTB_FWD: CACCATTGGCAATGAGCGGTTC

ACTB_REV: AGGTCTTTGCGGATGTCCACGT

Relative expression was calculated using the $\Delta\Delta C_t$ method with ACTB for normalization between samples.

## 2.5 Acknowledgement

We thank I. Fung and B. Wong for help with illustrations.

**Financial support**: Samsung Scholarship (to E.K.), Susan G. Komen Postdoctoral Fellowship PDF12230602 (to N.I.), Long-term postdoctoral fellowship by the European Molecular Biology Laboratory (to A.K.), Conquer Cancer Foundation of ASCO Young Investigator Award (to S.M.C), U.S. NCI grant, U01 CA176058 (to W.C.H.). The work was conducted as part of the Slim Initiative for Genomic Medicine, a project funded by the Carlos Slim Foundation in Mexico.

# Chapter 3

# Gene expression correlation analysis differentiates functional alleles from neutral alleles

This chapter is adapted from:

**Systematic functional interrogation of rare cancer variants identifies oncogenic alleles**

Cancer Discov. 2016 May 4. pii: CD-16-0160. [Epub ahead of print] PMID: 27147599

Eejung Kim*, Nina Ilic*, Yashaswi Shrestha*, Lihua Zou*, Atanas Kamburov*, Cong Zhu, Xiaoping Yang, Rakela Lubonja, Nancy Tran, Cindy Nguyen, Michael S. Lawrence, Federica Piccioni, Mukta Bagul, John G. Doench, Candace R. Chouinard, Xiaoyun Wu, Larson Hogstrom, Ted Natoli, Pablo Tamayo, Heiko Horn, Steven M. Corsello, Kasper Lage, David E. Root, Aravind Subramanian, Todd R. Golub, Gad Getz, Jesse S. Boehm, William C. Hahn

(*co-first author)

**Contribution:**

E.K., N.I., Y.S, G.G., J.S.B. and W.C.H. conceived the study.

M.S.L., and A.K. led cancer genomics analysis for variant selection.

Y.S., C.Z., X.Y., R.L., N.T., C.N. and D.E.R. generated cDNA reagents.

E.K., N.I., Y.S., C.R.C. and J.G.D. performed tumorigenesis experiments

F.P., M.B., X.W, Y.S. and D E.R. performed gene expression experiments.

E.K., L.Z, L.H., T.N. and S.M.C. performed gene expression analyses.

H.H., K.L., P.T. and S.M.C. advised on analysis strategy.

E.K., N.I., A.K., Y.S. and C.R.C. performed validation experiments.

E.K., J.S.B. and W.C.H. wrote the manuscript.

All authors participated on discussion and commented on the manuscript.

**Specific contribution:**

Eejung Kim generated all figures in this chapter.

## 3.1 Introduction

In the previous chapter, high throughput *in vivo* tumorigenesis screening was used to identify novel transforming alleles among 474 mutant alleles from significantly mutated genes curated from 5,338 tumors. Although this method is powerful in identifying rare transforming alleles such as *KRAS*$^{D33E}$, the numerous limitations discussed in section 2.4, make negative results uninterpretable.

We utilized the L1000 gene expression assay as a generalizable, function-agnostic method to interrogate these mutant alleles as well as their wild type counterparts and reference alleles of known biological functions. As discussed in 1.4.5, gene expression could be a functional readout of a cellular state that can be used to infer novel relationships between genetic perturbation and pharmacologic treatments (114). This approach complements the *in vivo* tumorigenesis screen to identify novel functional alleles.

## 3.2 Results

### 3.2.1 Gene expression correlation analysis differentiates allele function

In parallel to testing the tumorigenic potential of each allele *in vivo*, we created expression signatures for each of these alleles by expressing the 1163 constructs in a genetically defined, immortalized human embryonic kidney cell line (HA1E) (136). We selected this cell line since established cancer cell lines harbor many genetic alterations, which could confound the interpretation of expressing each allele. We decided to use HA1E cells, and not HA1E-M cells, which was used in the *in vivo* screen, because we wished to eliminate the contribution of the MEK$^{DD}$ allele. We measured transcript levels of 978 landmark genes using the L1000 Luminex bead-based gene expression assay (115) (Materials and Methods). Using the normalized gene expression change induced by each overexpressed allele, we calculated

the pairwise Spearman correlation coefficient of all the alleles included in the study (**Figure 3-1**). We excluded alleles with low infection efficiency (less than 40%), allowing us to assess 1036 perturbations (Materials and Methods; Supplementary Table S5).

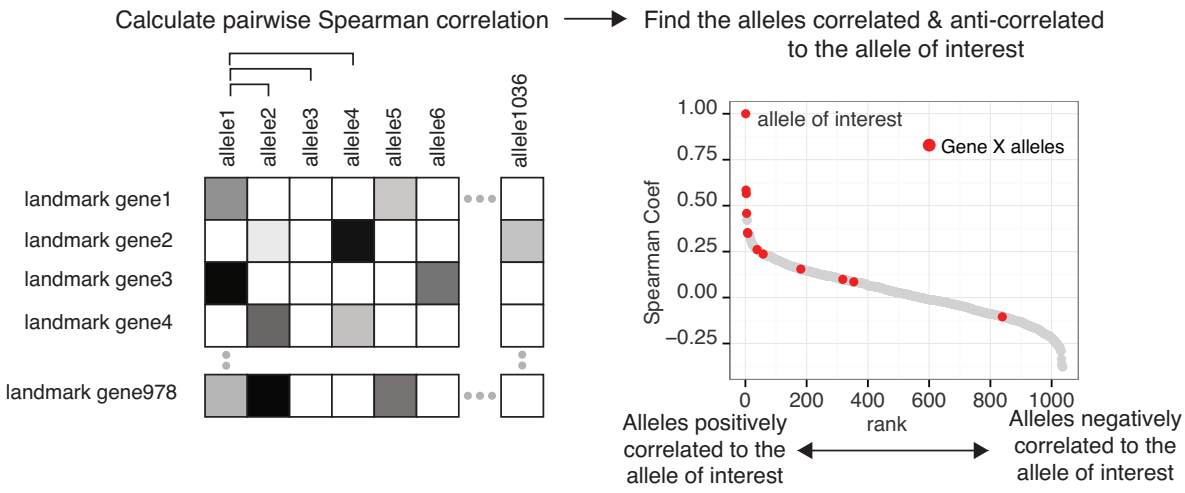



**Figure 3-1. Gene expression correlation analysis**

Expression signatures were analyzed by pairwise Spearman correlation to identify similar or dissimilar alleles to the allele of interest

Using the pairwise Spearman correlation coefficient between every pair of alleles included in the study, we first examined whether known relationships were detected. For instance, we found that the expression relationship of *KRAS*$^{G12V}$, a well-known gain-of-function mutant of *KRAS,* correlated highly with other known oncogenic *KRAS* and *NRAS* mutants (**Figure 3-2A**). Other known oncogenic alleles such as *AKT*$^{E17K}$ did not correlate with the *KRAS* signature, demonstrating that this correlation was not simply the consequence of a pro-survival signal induced by an oncogenic allele. Novel alleles of *KRAS*, D33E and E62K correlated less strongly to known *KRAS* activating mutants but were clearly differentiated from the wild type alleles, suggesting they may be activating mutants (**Figure 3-2A**). In addition, when we examined *NRAS*$^{Q61H}$, known activating mutant of *NRAS*, we found that this allele was highly

correlated with other oncogenic *NRAS* mutants but that the novel Y64D allele was more similar to wild type *NRAS* allele suggesting that this allele is likely to be a passenger allele (**Figure 3-2B**). Indeed, Y64D did not score in pooled *in vivo* screen.

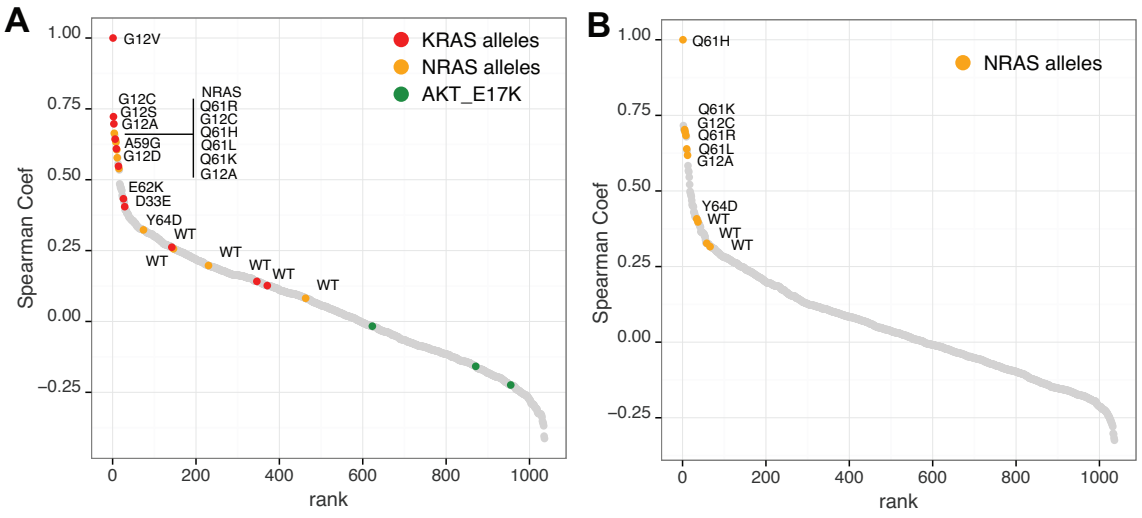



**Figure 3-2. Gene expression correlation analysis**

**(A)** *KRAS*$^{G12V}$ induces similar gene expression changes as other known activating alleles of *KRAS* and *NRAS*.

**(B)** *NRAS*$^{Q61H}$ induces similar gene expression changes as other known activating alleles of *NRAS*. However, the signature from the novel Y64D allele had a lower correlation, similar to wild type.

## 3.2.2 Gene expression allows differentiation between functional and neutral variants

The pattern of gain of function mutants showing higher correlation to other similarly activating mutants was also observed in other known oncogenes such as *IDH1/2* (**Figure 3-3**). We found that other known gain of function mutants *IDH2*$^{R172M}$, *IDH1*$^{R132C}$, *IDH1*$^{R132S}$, *IDH1*$^{R132H}$ and *IDH1*$^{R132L}$ were highly correlated to known gain of function mutant *IDH2*$^{R172K}$ (151). On the

other hand, the *IDH1* E190K and P33S alleles and the *IDH2* G137E, E268D, A416V, A47V, T331M, and I138F alleles failed to correlate to known activating mutants, suggesting these alleles were more similar to the WT allele (**Figure 3-3**).

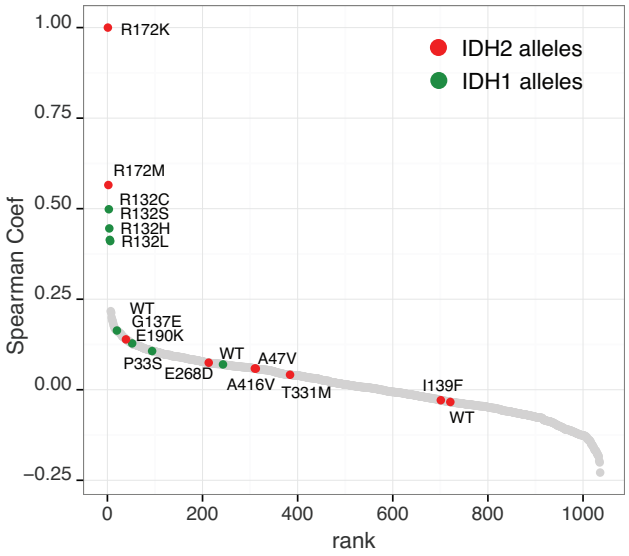



**Figure 3-3 Gain-of-function mutants of *IDH1* and *IHD2* are highly correlated.**

*IDH1/2* alleles were correlated to known activating mutant *IDH2*$^{R172K}$. Other known activating alleles of *IDH1/2* are highly correlated to *IDH2*$^{R172K}$.

Next, we investigated *PTEN*, a commonly mutated tumor suppressor gene, whose loss of function leads to constitutive activation of the phosphatidylinositol-3-kinase (PI3K) signaling pathway (152). Among the eight *PTEN* alleles included in this study, F90S, R233Q, K6N, and R173H correlated with the signature induced by overexpressing wild type *PTEN*, suggesting that these alleles did not completely inactivate PTEN function (**Figure 3-4A**). F90S mutant was recently shown to retain lipid phosphatase activity, but to be unable to translocate to plasma membrane (85). R233Q may also affect localization (153). R173H variant was previously

reported to lose phosphoinositide phosphatase activity (83), but its effect was later reported to be less severe than that of nonsense mutation (154). Our data supports that R173H retains residual PTEN function. In contrast, a known loss-of-function, dominant interfering allele (G129E) (82,84) failed to correlate with the wild type allele. We also found that signatures from the G129V, G127V and G127R alleles were clearly distinct from the wild type allele and moderately correlated to G129E (**Figure 3-4B, C**), suggesting that these alleles are also likely to be loss-of-function variants. Other alleles that activate PI3K signaling (*AKT1*$^{E17K}$) were anti-correlated with wild type *PTEN* (**Figure 3-4A**).

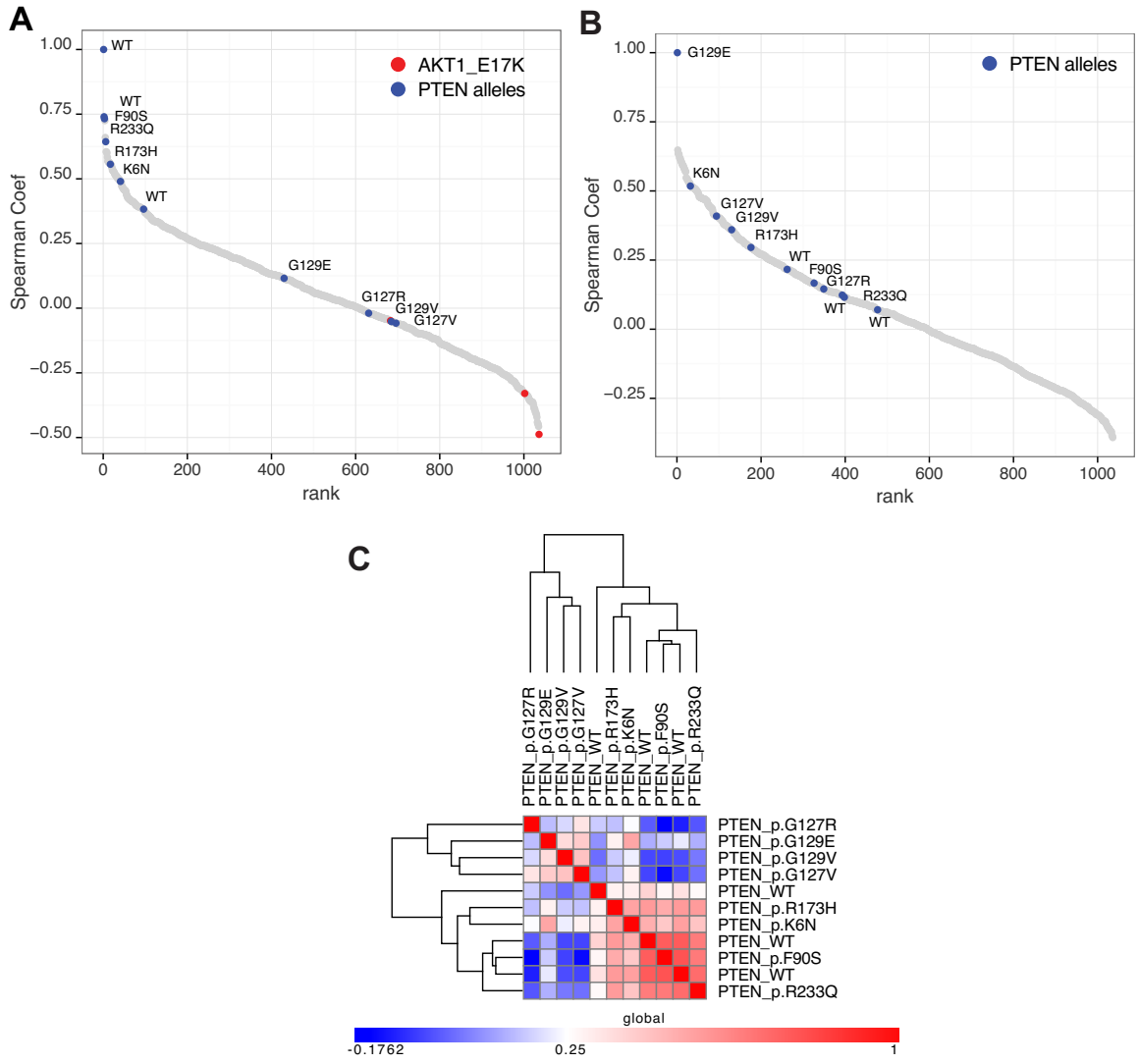



**Figure 3-4. Loss-of-function mutants of *PTEN* loose correlation to the wild type.**

(**A**) When correlated to the *PTEN* wild type, F90S, R233Q, K6N, R173H correlated strongly with the wild type *PTEN*. The known loss-of-function, dominant negative allele G129E showed a lower correlation. G127R, G129V, G127V also showed low correlation to the wild type.

(**B**) When alleles were correlated to *PTEN*$^{G129E}$, other likely loss-of-function alleles G12V, G129V, and G127R were only moderately correlated.

(**C**) When the gene expression changes induced by expression of *PTEN* allelic series were clustered, likely loss-of-function alleles were separated from the likely passenger mutants.

We used a similar approach to differentiate several alleles of *SPOP*, a gene mutated in prostate and endometrial cancers (155,156) (**Figure 3-5A**). Specifically, we found that the W131G, F133S, K134N, and W131C alleles strongly correlated with F102C, a known loss-of-function, dominant negative variant (157,158), but that the WT, K101I, E50K, and E47A did not correlate with the F102C allele. Codons F102, W131, F133 and K134 are mutated mostly in prostate cancers and E47 and E50 are altered in endometrial cancers (155,156,159). Recently, SPOP was shown to induce ubiquitination and degradation of androgen receptor and ERG in prostate cancer and estrogen receptor-alpha in endometrial cancer, but the SPOP mutants associated with respective cancer were unable to do so (157,158,160,161). When we looked for alleles correlated to E50K, loss-of-function allele in endometrial cancer (160), E47A was highly correlated, implying that this allele may also be loss-of-function (**Figure 3-5B**). Gene expression signatures of E47 and E50 variants clustered with that of wild type but were distinct from F102, W131, F133 and K134 variants (**Figure 3-5C**). These findings suggest that gene expression analysis may allow nuanced interpretation of loss-of-function alleles that are associated with specific context. Since missense mutations in tumor suppressor genes tend to occur throughout their coding sequences, it is often difficult to differentiate functional from non-functional mutations by inspecting of the mutations or their frequency. Examining gene expression changes induced by these mutations may facilitate the classification of missense mutant alleles.

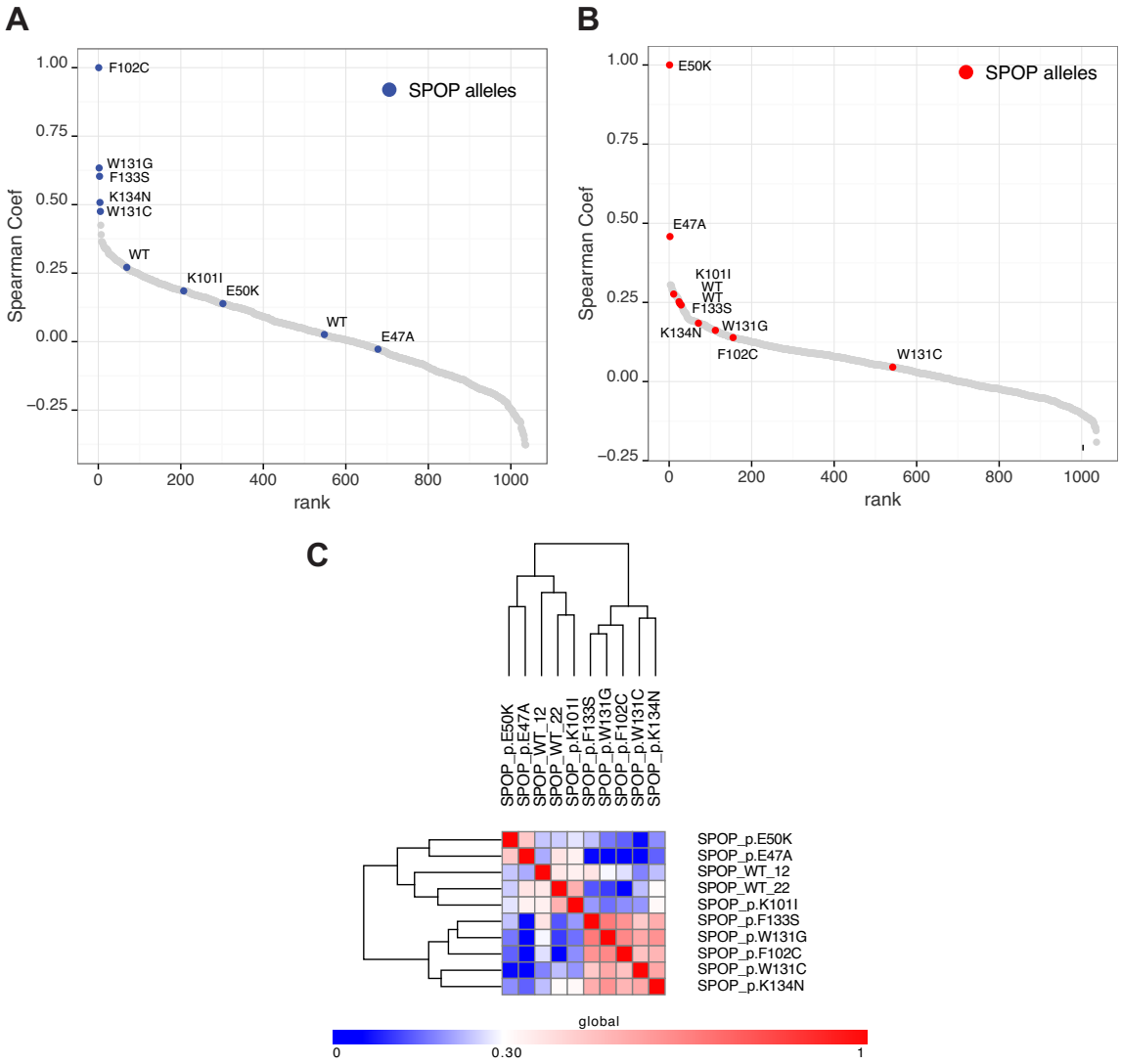



**Figure 3-5. Dominant negative alleles of *SPOP* are highly correlated.**

(**A**) When alleles were correlated against *SPOP*$^{F102C}$, a loss-of-function, dominant negative *SPOP* allele, other known loss-of-function, dominant negative alleles W131G, F133S, K134N, and W131C were highly correlated. On the other hand, E50K, K101I, E47A had lower correlation to F102C.

(**B**) When alleles were compared to *SPOP*$^{E50K}$, other likely loss-of-function allele E47A was highly correlated.

(**C**) When the gene expression changes induced by expression of *SPOP* allelic series were

**Figure 3-5. (Continued).**

clustered, likely loss-of-function, dominant negative alleles discovered in prostate cancer were separated from the wild type and likely loss-of-function alleles found in endometrial cancer.

### 3.2.3 Negative regulators of transcription factors are identifiable by gene expression analysis

We also examined which of the included alleles correlated with the proto-oncogene *MYC*, a commonly amplified oncogenic transcription factor (162). The most positively correlated allele in our dataset was wild type *BRD4*, which is a transcriptional activator of *MYC* (**Figure 3-6A**) (163). BRD4 has been shown to regulate *MYC* transcription, and pharmacologic modulation of BRD4 inhibited proliferation in MYC-dependent cancers (163). We found that the *FBXW7* wild type, R658Q, I347M, R689Q, and S462Y alleles were anti-correlated to wild type *MYC* (**Figure 3-6A**). FBXW7 is the substrate recognition component of the SCF ubiquitin ligase targeting MYC (164), suggesting that these four alleles do not affect FBXW7 function. In contrast, we found that the known dominant interfering alleles, *FBXW7* R505C, R465C, and R465H (165,166), were anti-correlated to wild type *FBXW7*, in consonance with the interpretation that these alleles inhibit endogenous wild type *FBXW7* (**Figure 3-6B**).

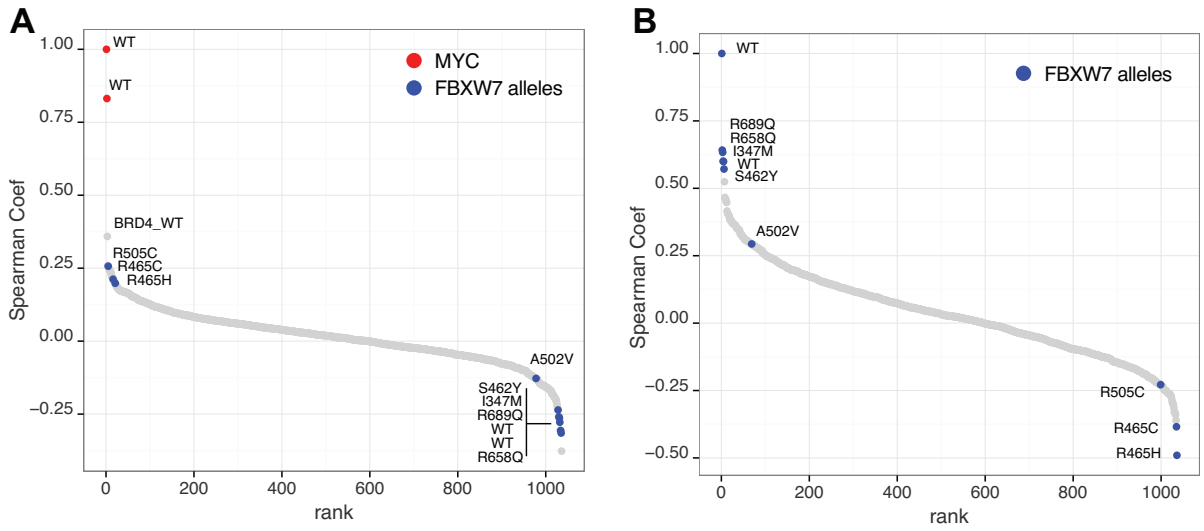



**Figure 3-6. *MYC* and *FBXW7* gene expression signatures are anti-correlated and dominant negative alleles of *FBXW7* are anti-correlated to that of wild type alleles.**

(**A**) *FBXW7* wild type, R658Q, I347M, S462Y, and R689Q, were strongly anti-correlated to MYC. Known dominant negative alleles (R505C, R465C, R465H) no longer were anti-correlated to *MYC*. *BRD4* wild type was the most closely correlated to MYC.

(**B**) When alleles were correlated to the *FBXW7* wild type, known dominant interfering alleles (R505C, R465C, R465H) were anti-correlated to the wild type.

Gene expression analysis of *NFE2L2* mutants showed a similar gene expression pattern to that of wild type, presumably because overexpression of the wild type allele may induce similar gene expression changes as does the overexpression of gain-of-function mutants in the short term gene expression assay (**Figure 3-7**). This is in contrast to the findings in the *in vivo* tumorigenesis in Chapter 2, where the gain-of-function mutants G31R, G31V, G31A and T80K exhibited robust tumor formation phenotype when compared to their wild type. These observations demonstrate that short term *in vitro* gene expression assays may not be able to differentiate overexpressed wild type and gain-of-function alleles.

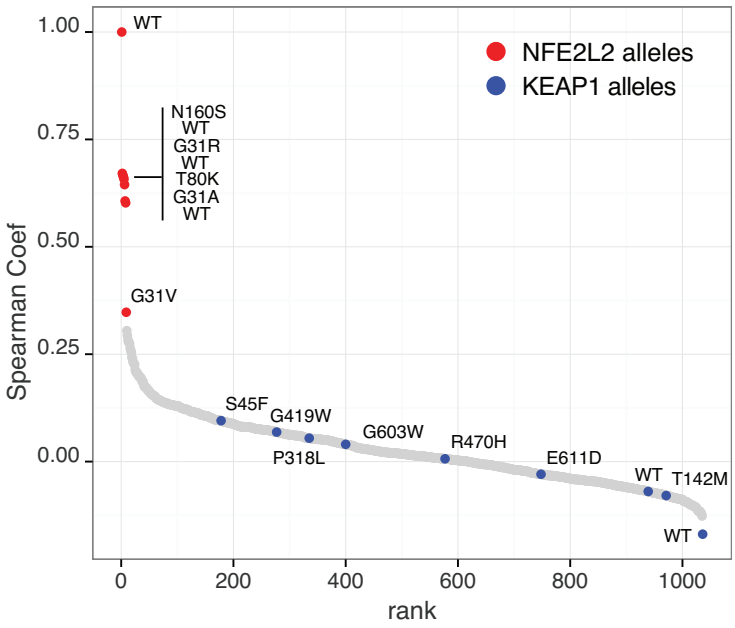



**Figure 3-7. NFE2L2 and KEAP1 gene expression signatures are anti-correlated.**

When alleles were compared to the *NFE2L2* wild type, all *NFE2L2* alleles were highly correlated to the wild type. *KEAP1* wild type alleles tended to anti-correlated with *NFE2L2* wild type.

### 3.2.4 Experimental characterization complements *in silico* method

To investigate whether high throughput functional phenotyping complements *in silico* predictions, we compared our observations pertaining to 71 alleles analyzed herein to four different *in silico* methods, Polyphen2 (56), Mutation Assessor (57), CHASM (58), and VEST (59). Each of these methods makes predictions about whether a mutation is likely to affect protein function but does not attempt to predict whether the mutation induces gain or loss of function. To compare these approaches, we used the term “functional variant” to denote both gain-of-function and loss-of-function alleles (61) and “neutral variant” for all other alleles. The concordance rates between each of these methods and our approach ranged from 66% to 77%

(Supplementary Table S6; **Figure 3-8**; Materials and Methods), suggesting that gene expression comparisons provided additional information about gene function. For example, Polyphen2 and CHASM predicted that *SPOP*$^{K134N}$ was likely to be a functional variant while Mutation Assessor and VEST assessed this to be a neutral allele. We found that *SPOP*$^{K134N}$ correlated with *SPOP*$^{F102C}$, providing evidence that this allele is a functional variant. Together, these observations suggest that the experimental characterization of alleles complements *in silico* methods.

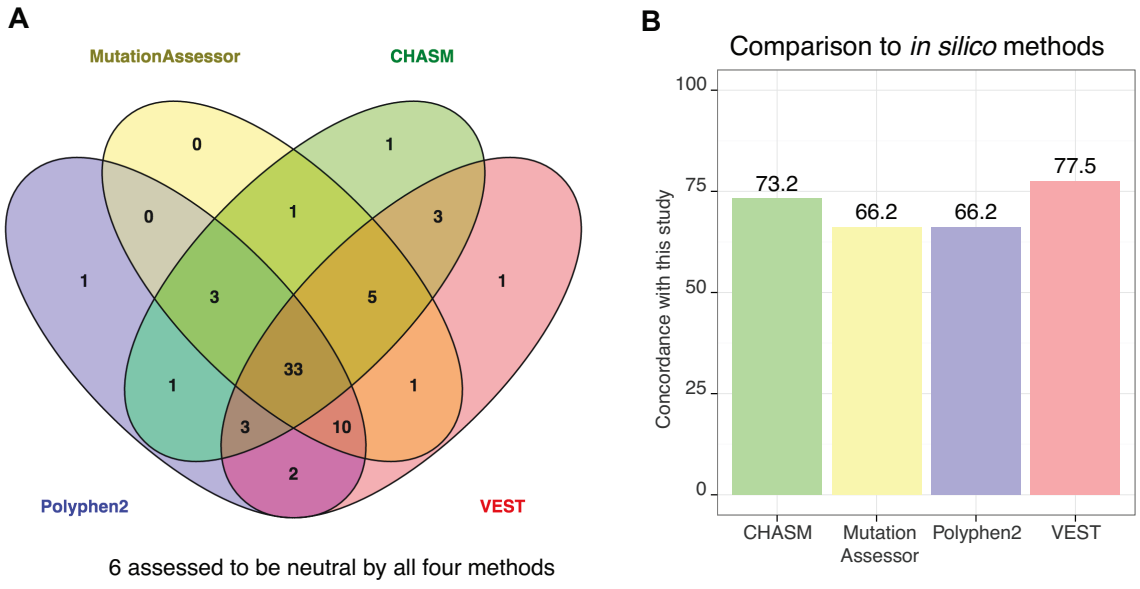



**Figure 3-8. Comparison to *in silico* methods.**

(**A**) Venn diagram of four different methods showing the overlap of the number of alleles called "functional" in each method. Please refer to Methods for description.

(**B**) Concordance rate of the four different *in silico* methods to the analysis from this study. The concordance rate ranged from 66 – 77%.

## 3.3 Discussion

Using gene expression signatures generated by expressing wild type or mutant alleles, we found that some *PTEN*, *FBXW7*, *NRAS, IDH1/2, and SPOP* alleles resembled the wild type alleles or known functional mutants, suggesting that these alleles are functionally similar to those alleles. On the other hand, in oncogenes such as *NFE2L2*, we found that gain-of-function mutants induced similar gene expression signatures as the wild type allele. This observation suggests that some truly transforming alleles may not score in the short term *in vitro* gene expression assay. Furthermore, for genes whose mechanism of action involves longer-term processes such as DNA repair, the acute effect of overexpressing alleles may not be reflected in gene expression changes. Combining expression profiling with tumor formation or other phenotypic experiments may provide complementary information in these cases.

In summary, results from Chapter 2 and Chapter 3 demonstrate that systematically performing functional assays complements the structural information gathered from the sequencing efforts to accelerate the interpretation of cancer associated variants. We anticipate that as additional tumors are characterized in both research and clinical settings, additional cancer associated genes and alleles will be identified, and the approach described here can be useful to ascertain the function of these alleles. Using diverse cellular backgrounds and different phenotypic assays will also increase the power to detect functional variants and reduce false negatives. As more functional data become available, we may also be able to gain insights on empirically improving the accuracy of mutation impact calling algorithms by incorporating information from high confidence functional data. This iterative process between functional and structural genomics will synergistically facilitate the complete description of cancer-associated mutations.

## 3.4 Materials and Methods

### Expression profiling using L1000

L1000 is a high-throughput, bead-based gene expression assay in which mRNA is extracted from cultured human cells treated with various chemical or genomic perturbagens (small molecules, gene knockdowns, or gene over-expression constructs). HA1E cells were plated at 400 cells per well in 384 well plates. The next day cells were transduced with 3 µl of lentiviral supernatant by spin infection. Infections were performed in 5 replicates, 2 of which were used to assess infection efficiency and the remaining 3 for gene expression profiling. Following 24-h incubation, media and virus were removed and replaced with complete growth media or media containing antibiotics (for infection efficiency calculation). Cell plates used for gene expression analysis were not selected to reduce the effect of antibiotics on the gene expression. 96 hours after infection, cells were lysed with addition of TCL buffer (Qiagen) and incubated for 30 minutes at room temperature. mRNA is reverse-transcribed into first-strand cDNA. Gene specific probes containing barcodes and universal primer sites are annealed to the first strand cDNA. The probes are ligated to form a template for PCR. The template is PCR amplified with biotinylated universal primers. The end products are biotinylated, fixed length, barcoded amplicons. The amplicons are then mixed with Luminex beads that contain complementary barcodes to those encoded in each of the 978 amplified landmark genes. These beads are then stained with fluorescent streptavidin-phycoerythrin (SAPE) and detected in 384 well plate format on a Luminex FlexMap flow cytometry-based scanner. The resulting readout is a measure of mean fluorescent intensity (MFI) for each landmark gene. The raw expression data are log2-scaled, quantile normalized, and z-scored, such that a differential expression value is achieved for each gene in each well. These differential expression values are collapsed across replicate wells using a weighted average to yield a differential expression signature for each perturbagen. Each replicate is weighted according to its correlation with the

others. These signatures were used for subsequent analysis. Detailed protocol is available at LINCS website (http://support.lincscloud.org/hc/en-us/categories/200155686-Data-Generation-Protocols).

**Gene expression correlation analysis**

Each normalized gene expression data was filtered by infection efficiency, which was calculated by dividing cell viability after antibiotic selection with cell viability without antibiotic selection by CellTiter-Glo Luminescent Cell Viability Assay (Promega). Viability was assessed 96h post-infection. 40% infection efficiency was used as cutoff to filter inadequately transduced alleles. 1036 gene expression signatures were Spearman correlated with gene expression signature of all other ORFs. “cor(method=”spearman”)” function in R was used for Spearman correlation coefficient calculation (167). Negative controls (BFP, eGFP, HcRed, LacZ, Luciferase), L1000 expression plate controls (NFE2L2, RHEB, NFKB1A, DNMT3A) were also included. After pairwise Spearman correlation, alleles at the extreme ends of the spectrum were manually curated to find alleles that are consistent with previously known relationship.

**Comparison to the *in silico* methods**

We compared our observations to four different in silico methods, Polyphen2 (56), Mutation Assessor (57), CHASM (58), and VEST (59). We used the term “functional variant,” to denote both gain and loss of function alleles (61), and “neutral variant” otherwise. For PolyPhen2, “possibly damaging” and “probably damaging” categories were considered functional. HumDiv-trained Polyphen2 was used. For Mutation Assessor, “high” and “medium” were considered functional. For CHASM and VEST, alleles with FDR <0.05 were considered functional. Default parameters were used for PolyPhen2 and Mutation Assessor and “cancer type: other” was chosen for CHASM analysis. The Venn diagram was drawn with Venny (168).

## 3.5 Acknowledgement

**Financial support**: Samsung Scholarship (to E.K.), Susan G. Komen Postdoctoral Fellowship PDF12230602 (to N.I.), Long-term postdoctoral fellowship by the European Molecular Biology Laboratory (to A.K.), Conquer Cancer Foundation of ASCO Young Investigator Award (to S.M.C), U.S. NCI grant, U01 CA176058 (to W.C.H.). The work was conducted as part of the Slim Initiative for Genomic Medicine, a project funded by the Carlos Slim Foundation in Mexico.

# Chapter 4

# Functional genomics approach to identify *FRS2* as amplified oncogene in high-grade serous ovarian cancer

This chapter is adapted from:

**The Tyrosine Kinase Adaptor Protein FRS2 is Oncogenic and Amplified in High-grade Serous Ovarian Cancer**

Mol Cancer Res. 2015 Mar;13(3):502-9.

Leo Y. Luo*, Eejung Kim*, Hiu Wing Cheung*, Barbara A. Weir, Gavin P. Dunn, Rhine R. Shen, William C. Hahn

(*co-first author)

**Contribution:**

L.Y.L. H.W.C, G.P.D and W.C.H conceived the study.

L.Y.L, E.K and G.P.D performed the experiments.

L.Y.L, E.K, H.W.C, G.P.D, B.A.W and W.C.H analyzed the data.

L.Y.L, E.K and W.C.H wrote the manuscript.

**Specific contribution:**

Leo Y. Luo generated Figures 4-1, 2, 3, 4A, 4B, 6 and 8

Eejung Kim generated Figures 4-4C, 5, and 7

## 4.1 Introduction

In previous chapters, I characterized the function of non-synonymous point mutations. In this chapter, I systematically identify and characterize the driver gene of a recurrent focal amplification in ovarian cancer.

Ovarian cancer is the second most common gynecologic malignancy and the most common cause of gynecologic cancer death in the United States (169). Histologically, ovarian epithelial carcinomas can be divided into high-grade serous, low-grade serous, endometroid, mucinous, and clear cell types. Clinically, high-grade serous ovarian cancer (HGSOC) accounts for 70-80% of all ovarian carcinomas and is characterized by its *de novo* invasive nature and initial sensitivity to platinum treatment. The molecular features of HGSOC include *BRCA1/2* and *TP53* mutations and widespread DNA copy number alterations (170). The lack of readily targetable mutations found in HGSOC has contributed to slow progress in developing molecularly targeted therapies for this subset of ovarian cancers.

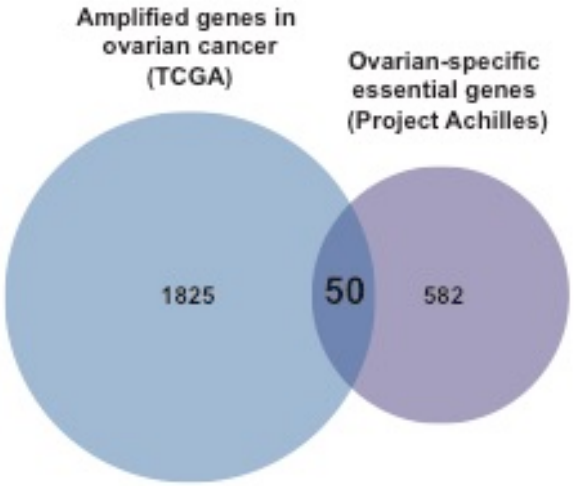
To catalog the molecular aberrations present in HGSOC, The Cancer Genome Atlas (TCGA) network performed a large-scale, multiplatform genomic profiling study of HGSOC (170). Analysis of 489 HGSOC primary tumors identified large number of recurrent somatic copy number alterations that include 31 focal amplifications. These amplified regions encode 1825 genes including known oncogenes such as *CCNE1* and *MYC*. However, the driver genes in the majority of the recurrently amplified regions remain unidentified.

In parallel to these genome characterization efforts, we initiated Project Achilles, a systematic effort to identify cancer dependencies at genome scale (171,172). Here by combining the output of ovarian cancer genome analysis with Project Achilles, we systematically interrogated 1825 recurrently amplified genes in ovarian cancer to identify genes that are essential in ovarian cancer cell lines that harbor such amplifications and identified *FRS2* as an amplified and essential gene in HGSOC.

## 4.2 Results

### 4.2.1 Identification of *FRS2* as an amplified and essential gene in ovarian cancer

High-grade serous ovarian cancers are characterized by high frequency, recurrent regions of copy number gain and loss. Recent genome-scale effort to characterize structural alterations in HGSOC has identified 31 recurrently amplified chromosomal regions containing total of 1,825 genes (170). To systematically study previously unknown lineage-specific dependencies, we initiated a genome-scale effort (Project Achilles) to identify genes essential for proliferation/survival of a large number of well characterized cancer cell lines using loss-of-function genetics with short hairpin RNAs (shRNA) (172). Although recent studies suggest that established ovarian cancer cell lines do not fully recapitulate the genetic alterations found in high grade ovarian cancers (173,174), here we have focused on those alterations found by the TCGA in human cancers and shared by these ovarian cancer cell lines. Using data from 102 cell lines of which 25 were from the ovarian lineage, we identified 582 ovarian-lineage specific gene dependencies (171). By looking at the intersection of genes involved in regions of recurrent copy number and essential in ovarian cancer cell lines, we identified 50 genes (**Figure 4-1**). Two of the 50 genes were previously identified as ovarian specific oncogenes (*PAX8*, *CCNE1*) using similar analytical approaches (171,175).




| | | | |
|---|---|---|---|
| 1 | ***PAX8*** | 26 | *MYST3* |
| 2 | *APOBEC2* | 27 | *OR2T1* |
| 3 | *CDK2* | 28 | *PABPC1* |
| 4 | *CXCL1* | 29 | *PABPC4* |
| 5 | *HRASLS* | 30 | *PEX13* |
| 6 | *NPC1* | 31 | *PNPT1* |
| 7 | *OR14A16* | 32 | *PRB3* |
| 8 | *PRKCQ* | 33 | ***PRKCE*** |
| 9 | *ABHD8* | 34 | *PTPRB* |
| 10 | *ADAMTSL4* | 35 | *RANBP9* |
| 11 | *AFP* | 36 | *RHAG* |
| 12 | *AGPS* | 37 | ***RPTOR*** |
| 13 | *AKR1CL1* | 38 | *SENP5* |
| 14 | *ALG8* | 39 | *SLC38A2* |
| 15 | *APOD* | 40 | *SLC45A2* |
| 16 | *C18orf45* | 41 | *SLC4A4* |
| 17 | ***CCNE1*** | 42 | *TACC3* |
| 18 | ***FRS2*** | 43 | *TBC1D16* |
| 19 | *GUCA1A* | 44 | *TFEB* |
| 20 | *HNRNPA3* | 45 | *TTR* |
| 21 | *HNRNPU* | 46 | *TXNDC5* |
| 22 | *LASS2* | 47 | *UQCR10* |
| 23 | *MRPL12* | 48 | *ZFP36L2* |
| 24 | *MTA3* | 49 | *ZNF398* |
| 25 | *MYLIP* | 50 | *ZNF623* |

**Figure 4-1. Amplified and essential genes in high grade serous ovarian cancer** (Leo Luo) *FRS2* is one of the 50 genes that are recurrently amplified in primary ovarian tumors and essential for ovarian cancer cell proliferation and survival.

Among the remaining genes, we focused on fibroblast growth receptor substrate 2 (*FRS2*) because FRS2 is (i) adaptor protein in the Fibroblast Growth Factor Receptor (FGFR) pathway, (ii) is located on chromosomal region 12q15, which is focally amplified in 12.5% of 559 primary high-grade serous ovarian cancers characterized by TCGA (**Figure 4-2**), and (iii) was among the top 100 genes that scored by our analysis of Project Achilles and copy number data in HGSOC.

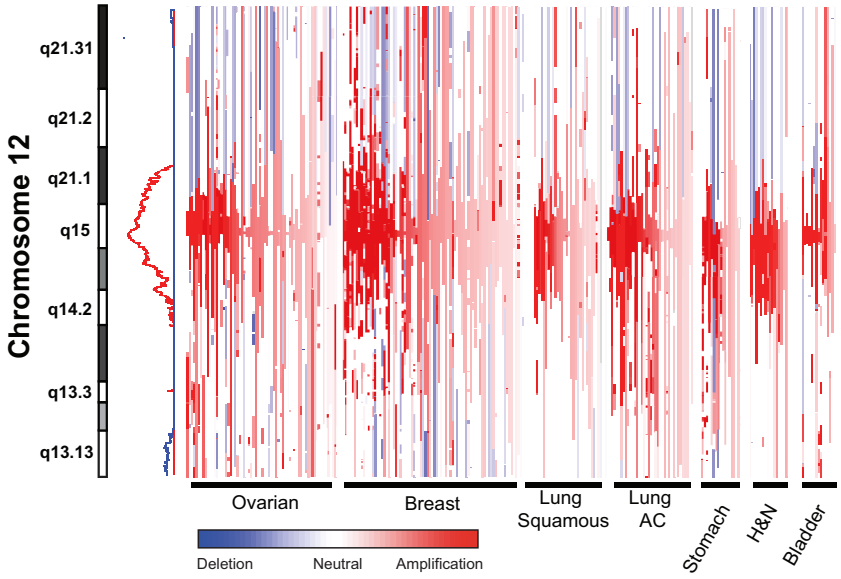



**Figure 4-2. FRS2 is located at the peak of amplified chromosomal region 12q15** (Leo Luo) Copy number profile along chromosome 12q of human tumor samples exhibits high level of *FRS2* amplification in multiple cancer types including ovarian, breast, lung squamous, lung adenocarcinoma, stomach, head and neck (H&N), and bladder. Each vertical line represents one tumor sample. Red is copy number gain, Blue is copy number loss.

We also found a structurally similar chromosomal region amplification in other cancer types such as breast invasive carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, head and neck squamous cell carcinoma, gastric adenocarcinoma, and bladder urothelial carcinoma. We used Genomic Identification of Significant Targets in Cancer Version 2.0 (GISTIC 2.0) algorithm to identify the peak of amplification, which corresponds to the highest level of copy number gain. In ovarian cancer samples, we observed the overlap between the peak of amplification and the location of *FRS2* gene. Furthermore, the focal amplification of 12q15 region in HGSOC is correlated with increased mRNA expression of *FRS2*, suggesting the functional relevance of the copy number gain (**Figure 4-3A**). In addition, we also observed frequent amplification of *FGFR* family of tyrosine kinase receptor genes in HGSOC. Strikingly, HGSOC samples that harbor 12q15 amplifications were often mutually exclusive with HGSOC

that harbor *FGFR1*, *FGFR2*, FGFR3, and *FGFR4* amplifications (Fisher's exact test *P*=0.028) (**Figure 4-3B**). This pattern of mutations is observed in commonly mutated genes in the same pathway, such as *KRAS* and *EGFR* mutations or *TP53* and *MDM2* mutations. These observations implicate FGF signaling through amplifications of FGFRs and FRS2 as a common event in HGSOCs.

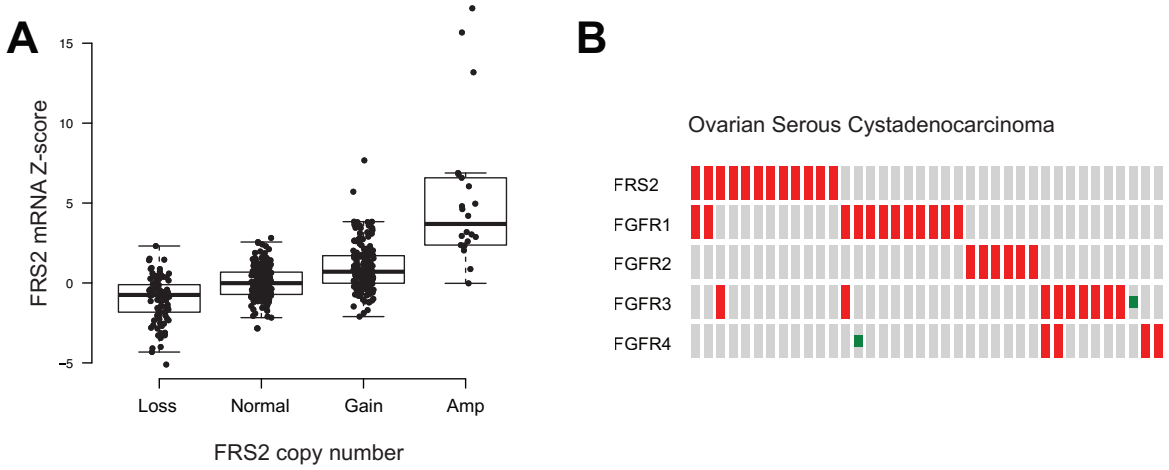



**Figure 4-3. FRS2 is located at the peak of amplified chromosomal region 12q15** (Leo Luo) (**A**) Level of *FRS2* mRNA expression in primary tumors correlates with the copy number. Copy number is divided into 4 categories based on log2 of copy numbers. "Amplification" is defined as Log2(Copy number) more than 1; "Gain" is between 0.2 and 1; "Normal" is between -0.2 and 0.2; "Loss" is less than -0.2.

(**B**) *FRS2* amplification and *FGFR1*, *FGFR2*, *FGFR3*, and *FGFR4* amplifications are mutually exclusive in high-grade serous ovarian cancers.

### 4.2.2 *FRS2* is essential in cancer cell lines that harbor 12q15 amplification

To confirm that *FRS2* was essential in *FRS2* amplified cancer cell lines, we used two independent shRNAs to suppress FRS2 expression in three cell lines with 12q15 amplification (CAL120_BREAST, COV644_OVARY, HCC1143_BREAST) and three cancer cell lines that

contain normal copies of 12q15 (CAOV3_OVARY, EFO21_OVARY, COV362_OVARY). We used both breast and ovarian cancer cell lines since we found focal amplification of 12q15 in a large subset of the primary breast cancers (**Figure 4-2**). Copy number data for these cell lines were obtained from the Broad Institute/Novartis Cancer Cell Line Encyclopedia (CCLE) (176) (**Figure 4-4A**). We found that FRS2 suppression by two independent shRNAs significantly decreased the proliferation of cancer cell lines that harbor the 12q15 amplification, when compared to cells that exhibit diploid copy number at 12q15 or cells infected with control shRNA (**Figure 4-4B**). The degree of FRS2 suppression in 12q15 amplified cell lines was validated by quantitative real-time PCR (**Figure 4-4C**). To demonstrate that FRS2 suppression induced apoptotic cell death in 12q15 amplified cell lines, we interrogated poly ADP-ribose polymerase (PARP) cleavage after suppression of FRS2 and sub-G1 fraction by flow cytometry. We found increased level of cleaved PARP in 12q15 amplified cell lines compared to cell lines without 12q15 amplification (**Figure 4-5A**). Similarly, we observed increased sub-G1 fraction upon suppression of FRS2 in *FRS2* amplified cell lines (**Figure 4-5B**). Together, these findings demonstrate that cancer cells that harbor 12q15 amplification require FRS2 expression for proliferation and survival.

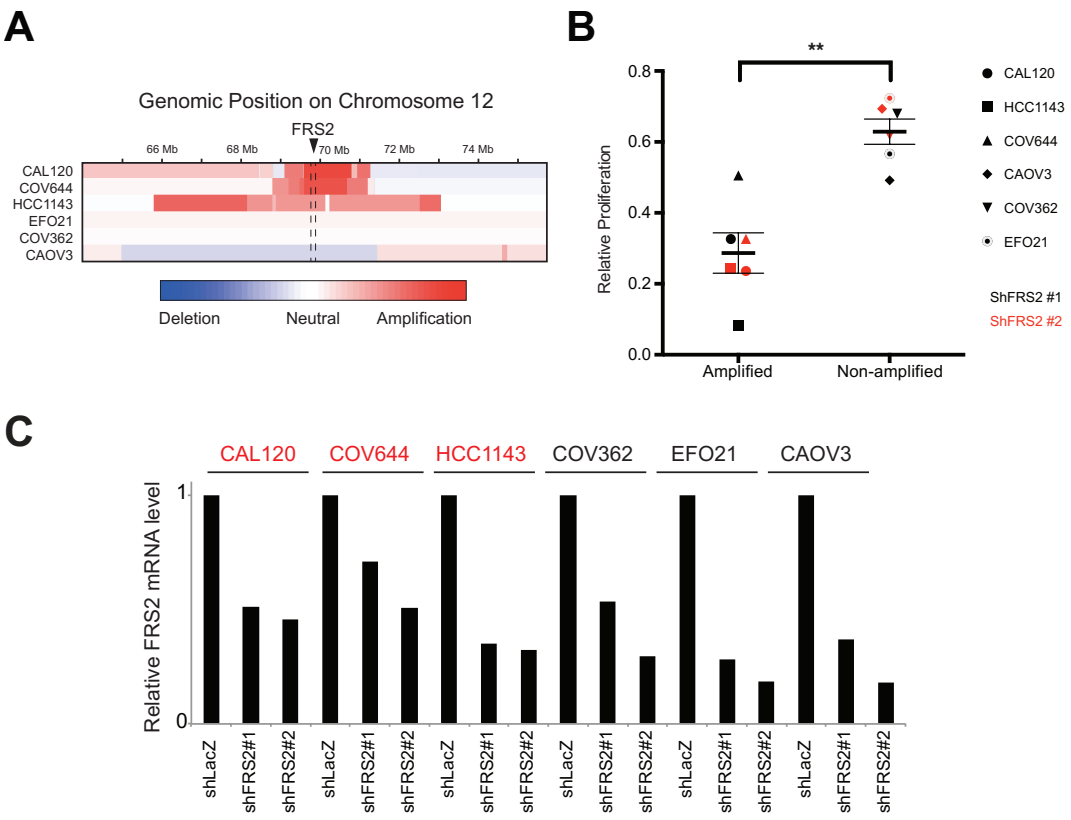



**Figure 4-4. Suppression of *FRS2* decreases the proliferation of ovarian and breast cancer cells harboring 12q15 amplification.**

**(A)** SNP array colorgram showing genomic amplification of chromosome 12q15 in ovarian and breast cancer cell lines. Red indicates copy number amplification and blue indicates copy number deletion. (Leo Luo)

**(B)** Proliferation effect of *FRS2* suppression on cancer cell lines that either harbor 12q15 amplification (CAL120, HCC1143, COV644) or normal copy number of 12q15 (CAOV3, COV362, EFO21) normalized to cells treated with shLacZ. Red: cell lines treated shFRS2 #1. Black: cell lines treated with shFRS2#2. ***P* < 0.01 compared to control shLacZ, Student's *t* test was used. (Leo Luo)

**(C)** Quantitative RT-PCR of FRS2 expression in FRS2 amplified (red) and non-amplified (black) cell lines.

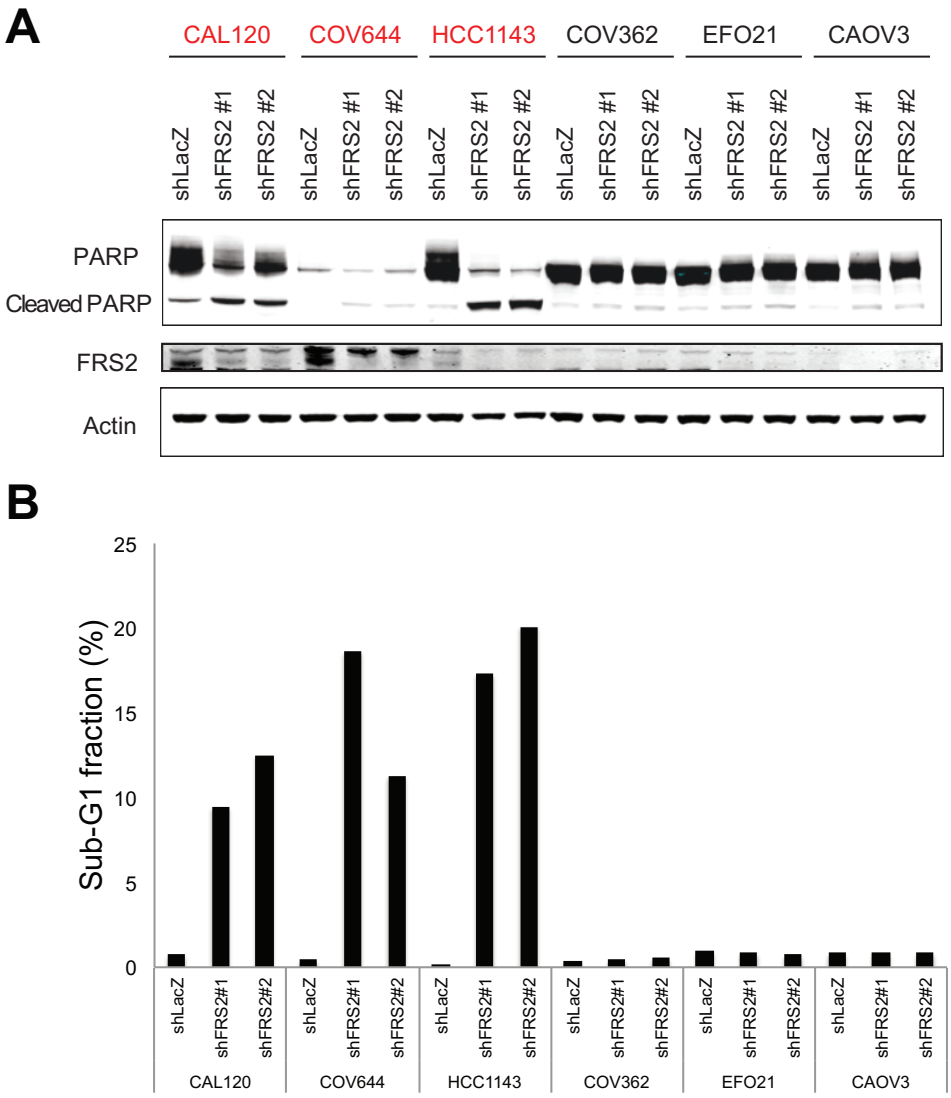



**Figure 4-5. Suppression of *FRS2* increases apoptosis in ovarian and breast cancer cells harboring 12q15 amplification.**

**(A)** Increased apoptosis in FRS2 amplified cell lines (red) upon FRS2 suppression, shown by increased PARP cleavage.

**(B)** Increased apoptosis in FRS2 amplified cell lines (red) upon FRS2 suppression, shown by increased sub G1 fraction population by flow cytometry.

### 4.2.3 FRS2 induces oncogenic transformation

To determine if FRS2 can contribute to tumorigenesis by inducing transformation, we performed anchorage-independent growth assays and tumor xenograft experiments. In our prior studies, we have shown that human kidney epithelial cells are immortalized by co-expression of the human catalytic subunit of telomerase (hTERT) and the SV40 Early Region (HA1E cell) and the expression of oncogenic alleles of RAS confers the ability to grow in anchorage-independent manner (136). We also demonstrated the *RAS* oncogene can be replaced by combination of downstream effectors of RAS signaling pathway, such as constitutively-activated MEK1 (MEK-DD) and AKT1 (myristoylated AKT) (21). In addition, we used the same genetic elements to immortalize human ovarian epithelial cells (IOSE) and used this cell line to identify ovarian cancer oncogenes such as *ID4* (177). The origin of HGSOG is still controversial as there are evidences supporting fallopian tube and ovarian surface epithelium hypotheses, but in our hands, there was no difference in transformation outcome in either model (137,177-180).

As previous studies have shown that FRS2 preferentially activates MAPK pathway, we overexpressed FRS2 in HA1E cell lines expressing constitutively active myristoylated AKT (HA1E-A) to determine whether FRS2-mediated MAPK pathway activation complemented AKT pathway activation to induce transformation. We measured anchorage independent growth with FRS2 overexpression and found that FRS2 overexpression was sufficient to induce anchorage independent colony formation of HA1E-A cells compared to cells expressing the control LacZ (**Figure 4-6A**). The number of colonies formed with FRS2 overexpression is significantly higher ($P<0.001$) compared to constitutively activated MEK, suggesting possible activation of additional pathways that contribute to the transformation process. We also conducted the same experiment in IOSE cells to show that FRS2 also induced transformation in ovarian epithelial cells (**Figure 4-6B**).

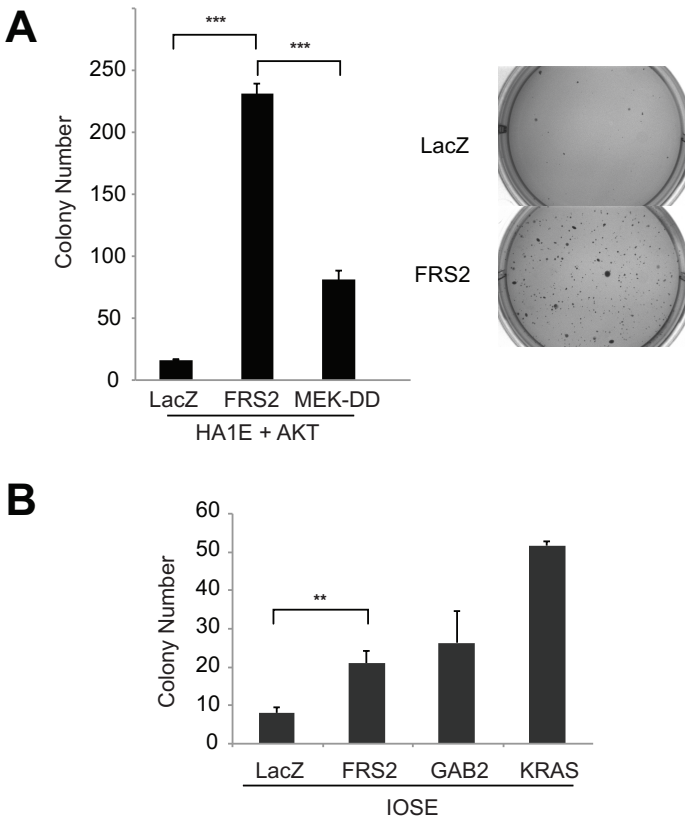



**Figure 4-6. FRS2 overexpression transforms immortalized cell lines.** (Leo Luo)

**(A)** *FRS2* promotes anchorage-independent growth in HA1E-A cells compared to LacZ control. MEK-DD, a constitutively active MEK, is positive control. Right, images of soft agar colonies formed by HA1E-A with either *FRS2* or control vector overexpression.

**(B)** *FRS2* promotes anchorage independent growth of IOSE (immortalized human ovarian epithelial) cells. GAB2 is a similar adaptor protein known to transform ovarian epithelial cells. $^{**}P < 0.01$, $^{***}P < 0.001$ compared to respective control vectors, Student's *t* test.

Next, we determined whether expression of FRS2 also induced tumor formation *in vivo* by expressing FRS2 in NIH/3T3 mouse fibroblast cells and implanting these cells subcutaneously in immunodeficient mice. We also conducted the same experiment in HA1E-A, but the result was inconclusive due to high background. At 11 weeks, we observed that tumors

formed in 33% (2 out of 6) of the injection sites harboring cells expressing FRS2 but failed to observe any tumors in sites harboring control cells (**Figure 4-7**). We note that since we implanted tumors in several sites in each mouse, and we terminated the experiment prior to observing tumor growth in all sites, these experiments may underestimate the tumorigenicity of these cells. These observations confirm that FRS2 overexpression can induce oncogenic transformation in human kidney fibroblasts or mouse fibroblasts by promoting anchorage-independent growth or in vivo xenograft tumor formation.

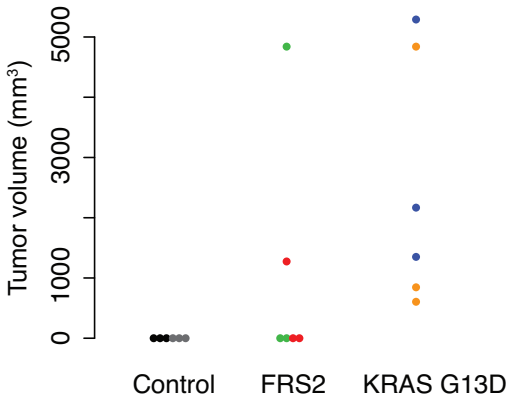



**Figure 4-7. FRS2 overexpression promotes tumorigensis *in vivo*.**

*FRS2* overexpression in NIH/3T3 promotes *in vivo* tumorigenesis in immunocompromised mice. Tumors from the same mouse are colored with the same color.

## 4.2.4 FRS2 amplification activates the MAPK pathway

Previous studies have shown that FRS2 is a critical mediator of FGFR signaling and plays an important role in activating MAPK and PI3K pathways (**Figure 4-8A**)(181,182). We have confirmed FRS2 overexpression induces activation of MAPK pathway in 293 HEK cells and IOSE cells by assessing phospho-Thr202/Tyr204 ERK1/2 levels (**Figure 4-8B**). Conversely,

suppression of FRS2 in FRS2-amplified cancer cell line caused a decrease in phospho-ERK levels (**Figure 4-8C**). In contrast, we failed to observe a change in phospho-AKT when we overexpressed FRS2 (**Figure 4-8B**). These observations suggest that FRS2 overexpression preferentially activates MAPK pathway in this context. This finding corroborates the results of anchorage-independent growth assays where we observed that FRS2 was able to induce increased colony growth when expressed with Myr-AKT as compared to co-expression with MEK-DD in HEK cells.

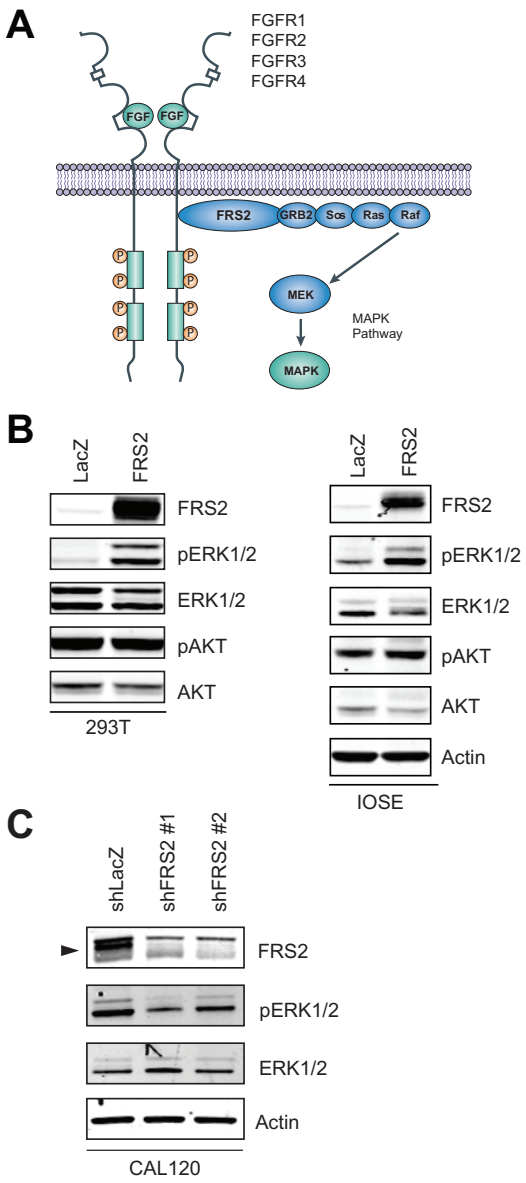



**Figure 4-8. *FRS2* promotes tumorigenesis via activation of MAPK pathway.** (Leo Luo)

**(A)** FRS2 functions as an adaptor protein in the fibroblast growth factor receptor signaling pathway, adapted from Turner and Grose (183).

**(B)** Effect of *FRS2* overexpression on phosphorylation of ERK in 293T cells and ovarian epithelial cells.

**(C)** Effect of *FRS2* suppression on phosphorylation of ERK in cancer cell line with 12q15 amplification.

## 4.3 Discussion

Here we identified *FRS2* as one of the 50 genes that are recurrently amplified in high-grade serous ovarian cancers (HGSOC) and essential to survival in ovarian cancer cell lines. *FRS2* belongs to the 12q15 genomic region that is focally amplified in 12.5% of HGSOC. Using independent shRNAs targeting against *FRS2*, we showed the expression of *FRS2* was essential for survival in cancer cells with 12q15 amplification. We also discovered that overexpression of *FRS2* in immortalized kidney fibroblast or ovarian epithelial cells promoted anchorage independent growth and tumorigenesis in mice. Together these observations nominate FRS2 as an amplified oncogene in a subset of high-grade serous ovarian cancers.

In addition to HGSOC, 12q15 amplification containing *FRS2* is found in other cancer types. 12q15 amplicon containing *FRS2* is focally amplified in 9.2% of breast invasive carcinomas. Indeed, we found that breast cancer cell lines that harbor 12q15 amplification are also sensitive to suppression of FRS2. Furthermore, new evidence has suggested the oncogenic role of *FRS2* and 12q15 amplification in high-grade liposarcomas through whole-exome sequencing and demonstrated sensitivity of FRS2-amplified high-grade liposarcoma cell lines to FRS2 suppression through shRNAs (184,185). These studies support *FRS2* as a *bona fide* oncogene in a variety of cancers and a potential therapeutic target for a subset of cancers that harbor such amplification.

The discovery of *FRS2* as an amplified oncogene adds to the family of FGFR signaling components that are critical to tumorigenesis in many cancer types. It is known that mutations or amplifications of *FGFR*s are frequent and inhibitor-sensitive in bladder cancer (186), gastric cancer (187), endometrial cancer (188), and non-small cell lung cancers (189,190). Large-scale genome-wide association studies have also linked breast cancer risk loci to *FGFR2* (191). More recent studies revealed the importance of FGF ligands, such as FGF19 amplification in liver cancer (192) and the therapeutic effect of neutralizing anti-FGF antibodies (193).

The 12q15 genomic region contains 15 genes with *FRS2* residing at the copy number peak of the amplicon (Supplementary Table S7). Prior work in high-grade liposarcoma, which has a broader region of amplification (12q13-12q15) than HGSOC, has suggested in addition to *FRS2*, other genes such as *CDK4* and *MDM2* may be driving events (184). Although neither *CDK4* nor *MDM2* is located within the 12q15 amplified region in HGSOC, we do not preclude the possibility that other genes in the genomic region may cooperate to drive various stages of tumorigenesis. Indeed, we recently demonstrated that multiple genes resident in a recurrently amplified region (3q26) contribute to cell transformation by inducing different cancer associated phenotypes, suggesting that further studies involving other assays will be necessary to investigate the function of these other genes (194).

Here we show a new functional class of adaptor proteins as driver oncogene in ovarian cancer. The adaptor proteins lack intrinsic enzymatic activities but mediate protein-protein interactions that drive protein complex formation. Classic examples of adaptor proteins include GRB2 in receptor tyrosine kinase signaling (195) and MYD88 in NF-κB signaling (196). FRS2 was originally discovered as a docking site for coordinated assembly of a multi-protein complex that include GRB2, GAB1, and SOS1, and serves a critical role in the FGFR signaling pathway (**Figure 4-8A**)(181,197). Unlike the signaling-amplifying activity of kinases, adaptor proteins are bottlenecks of the signaling pathway due to their stoichiometric relationship with interacting partners. Therefore, amplification or overexpression of the adaptor proteins can significantly alter the flux of the signal, thus carry important therapeutic implications such as mediating resistance to targeted therapy against receptor tyrosine kinase or conferring de novo sensitivity to signaling pathway inhibitors. Our laboratory has previously identified *CRKL*, an adaptor protein involved in RAS and RAP signaling, as an amplified oncogene in NSCLC (198). It was demonstrated that *CRKL* overexpression can mediate resistance to EGFR inhibitor in *EGFR*-mutant lung cancer cells and its amplification has been observed in gefitinib-resistant lung tumors. More recently, through a multiplexed *in vivo* transformation screen, we found another

adaptor protein, GAB2, as an amplified ovarian cancer oncogene that activates PI3K signaling (199). Ovarian cancer cells with *GAB2* alteration are sensitive to PI3K-pathway inhibition. An independent analysis of TCGA datasets across 16 cancer types has generated 75 amplified genes with druggable properties, including *FRS2* and EGFR family adaptors *GRB2* and *GRB7* (200). These findings suggest a new class of targetable oncogenes that are sensitive to exhibiting RTK signaling pathway inhibitors and present a new therapeutic opportunity to those patients with such genetic alterations.

## 4.4 Materials and methods

### Analysis of TCGA primary tumor data

Regions of copy number amplification identified by Genomic Identification of Significant Targets in Cancer (GISTIC) analyses were used from the TCGA study on high-grade serous ovarian cancer (170). All RefSeq genes within these regions of amplification (n = 1825) were identified and cross-referenced with genes interrogated in the Achilles screening library (n = 582). All primary HGSOC data were downloaded from the TCGA portal (http://tcga-data.nci.nih.gov/tcga). Genomic characterization data were visualized using the Integrative Genome Browser (http://www.broadinstitute.org/igv). Mutual exclusivity analysis was performed using the cBio Portal for Cancer Genomics (201,202).

### Analysis of shRNA screening data

Data from genome-scale loss of function screening was processed as described (171). Briefly, 54,000 shRNAs were lentivirally delivered to 102 cancer cell lines and the degree of representation of each shRNAs in the final cell population was measured by custom Affymetrix array. Normalization, variance stabilization and expression score calculation were conducted as specified in modified dCHIP method (172). Scores were median-adjusted per cell lines. Ovarian-

specific gene dependencies were determined with three complementary methods: (i) 150 best single shRNA or (ii) 300 second best shRNA or (iii) composite of all shRNAs for the gene using KS statistics. 582 genes (5.2%) were selected from the union of three methods above.

To identify genes that were both amplified in ovarian tumors and essential in amplified cancer cell lines, each gene identified as amplified in primary ovarian tumors (1,825 genes) was tested across the entire panel of 102 cell lines screened. Only genes with more than 5 amplified cell lines were included in the study. Amplified genes that had mapped shRNAs with a $P < 0.05$ were identified as candidate genes.

**Cell culture and generation of stable cell lines**

All human cancer cell lines were cultured in previously described media supplemented with 10% fetal bovine serum (FBS, Sigma) (171). Immortalized human ovarian surface epithelial cells (IOSE) (203) were maintained in 1:1 Medium 199: DMEM supplemented with 10% FBS. CAL120, COV644, COV362, and CAOV3 cells were cultured in Dulbecco's modification of Eagle's medium (Invitrogen) with 10% FBS. HCC1143, EFO21 cells were cultured in RPMI-1640 medium (Invitrogen) with 10% FBS. NIH/3T3 cells were cultured in DMEM with 10% bovine calf serum. Lentiviruses were produced by transfection of 293T packaging cells with a three-plasmid system. To generate stable cell lines, cells were seeded into 6-well dishes for 24 h before infection with 0.3 ml of lentiviruses for 12 h in the presence of 8 μg/ml polybrene. After the incubation, medium was replaced with fresh medium for another 24 h before selection in media containing 2 μg/ml of puromycin or 10 μg/ml of blasticidin until the control cells were no longer viable.

**Plasmids**

Human *FRS2* (from the CCSB human ORFeome collection (204)) was cloned into pLenti6.3-blast (*BamHI* and *BsrGI* sites). The pLX304-LacZ was used as a control vector. The

human *MEKD218, D222* (or *MEKDD)* fragment was removed from pBabe-puro-MEKDD plasmid (21) with *BamHI* and *SalI* and inserted into pLX304-Blasticidin. Lentiviral pLKO.1-puro-shRNA constructs were obtained from The RNAi Consortium or designed by custom oligo synthesis (IDT). The shRNA constructs used are as follows: control shRNA targeting *LacZ* (TRCN0000231710), *FRS2-specific* shRNAs (shFRS2#1: TRCN0000370440, shFRS2#2: 5'-CTCTAAATGGCTACCATAATA-3')

**Cell proliferation assay**

CAL120, COV644, HCC1143, EFO21, CAOV3, and COV362 cells (3 X $10^3$) were seeded into each well of 96-well plates 24 h prior to infection. Six replicate infections were performed for control shRNAs and each gene-specific shRNA in the presence of 8 μg/ml polybrene for 24 h followed by selection with 2 μg/ml of puromycin. The ATP content was measured at 6 days post-infection by using CellTiter-Glo luminescent cell viability assay (Promega).

**Anchorage-independent growth assay**

Growth in soft agar was determined by plating $5\times10^4$ cells in triplicate in 4 ml of medium containing 0.35% Noble agar (BD Biosciences), which was placed on top of 4 ml of solidified 0.6% agar. Unstained colonies greater than 100 μm in diameter were counted 4 weeks after plating using Cell Profiler software (205).

**Immunoblotting**

Cell lysates were prepared by scraping cells in lysis buffer (50 mM Tris HCl (pH 8), 150 mM NaCl, 1% Nonidet P40, 0.5% sodium deoxycholate, and 0.1% SDS) containing complete protease inhibitors (Roche) and phosphatase inhibitors (10 mM Sodium Floride and 5 mM Sodium Orthovanadate). Protein concentration was measured by using BCA Protein Assay kit

(Pierce). An equal amount of protein (20 μg) was separated by NuPAGE Novex Bis-Tris 4-12% gradient gels (Invitrogen) and then transferred onto a polyvinylidene difluoride membrane (Amersham). Antibodies against FRS2 (sc-8318) were purchased from Santa Cruz Biotechnology. Antibodies for PARP (#9532), phospho-ERK1/2 (#9101), ERK1/2(#9102) were purchased from Cell Signaling Technology and antibody specific for β-actin was obtained from Santa Cruz (sc-8432-HRP).

After incubation with the appropriate HRP-linked secondary antibodies (Bio-Rad), signals were visualized by enhanced chemiluminescence plus Western blotting detection reagents (Amersham). Alternatively, membrane was incubated with IRDye fluorescent secondary antibodies (LI-COR) and visualized by Odyssey quantitative fluorescence imaging system (LI-COR).

**Real-time quantitative RT-PCR**

Total RNA was extracted with RNeasy mini kit (Qiagen). Reverse transcription was performed using SuperScript III First-Strand Synthesis System (Invitrogen). Quantitative RT-PCR reactions were performed using SYBR green PCR Master Mix (Applied Biosystems). The primer sequences used were obtained from MGH PrimerBank:

FRS2_forward: CTGTCCAGATAAAGACACTGTCC

FRS2_reverse: CACGTTTGCGGGTGTATAAAATC

GAPDH_forward: CCTGTTCGACAGTCAGCCG

GAPDH_reverse: CGACCAAATCCGTTGACTCC

Triplicate reactions for the gene of interest and the endogenous control (GAPDH) were performed separately on the same cDNA samples by using the ABI 7900HT real-time PCR instrument (Applied Biosystems). The mean cycle threshold (Ct) was used for the ddCt analysis method.

**Flow cytometry**

Cells were collected, washed, and fixed with 70% ethanol at -20C for 4 hours. Fixed cells were washed, re-hydrated and re-suspended in propidium iodide staining solution (25ug/ml propidium iodide, Sigma P4862, 50ug/ml RNase A, Invitrogen 12091-021, in PBS) at room temperature for 30 minutes. Flow cytometry was done on BD LSR II flow cytometer (BD Biosciences). Debris and aggregates were gated out and the sub-G1 population was analyzed using FlowJo software.

**Tumorigenicity assay**

Female NCR/nude mice (Charles River Laboratories) were obtained at 6 weeks of age. All animal experiments were approved by the Dana-Farber Institutional Animal Care and Use Committee. Tumor xenograft experiments were performed as described (21). NIH/3T3 cells expressing indicated constructs were trypsinized and collected in fresh media. Cells were washed and re-suspended in PBS at $10^6$ cells per 100 ul. Cells were injected subcutaneously on left and right flanks, and upper back. Two mice were used for each experimental condition. 2 X $10^6$ cells were injected per site, three sites per mouse. Tumor injection sites were monitored for 3 months for tumor formation. Mice were euthanized when the largest tumor on mouse reached 2 cm in largest dimension.

**Statistical analysis**

Unless otherwise indicated, one-way ANOVA was used (GraphPad). $P < 0.05$ was considered statistically significant. Fisher's exact test was used for tumor formation assays and mutual exclusivity analysis. Two-tailed Student's *t* test was used for pairwise comparisons. A log-rank test was performed for animal survival studies.

# Chapter 5

# Conclusion

In this thesis I describe methods to improve the functional characterization of focal amplifications and non-synonymous point mutations. Integrating functional and structural genomics data facilitates identification of driver genes in recurrently amplified regions. Indeed, investigating amplified and essential genes in high-grade serous ovarian cancer led to the discovery of *FRS2* as an oncogene. This method may allow identification of driver genes in other recurrently amplified regions in ovarian cancers. As CRISPR technology is increasingly utilized, genome-scale CRISPR screen data may provide additional information to pinpoint essential and amplified genes.

To study the functional impact of point mutations, 474 alleles, the majority of which was observed only once in a set of 5,338 tumors, were generated and subjected to two functional assays, a pooled *in vivo* tumorigenesis screen and gene expression profiling. I demonstrated that the large-scale experimental characterization of cancer-associated gene variants is feasible and can generate valuable insights. From the *in vivo* tumorigenesis screen, I found that rare variants could be driver events in tumorigenesis. Alleles such as *KRAS*$^{D33E}$, *POT1*$^{G76V}$ and *PIK3CB*$^{A1048V}$ were shown to be transforming. *NFE2L2* alleles were shown to be transforming *in vivo* for the first time as well. Through analyzing gene expression correlation to wild type and known functional alleles, I inferred the functional status of unstudied alleles in genes including *KRAS*, *NRAS*, *IDH1/2*, *SPOP*, *PTEN* and *FBXW7*. These methods provide proof-of-concept evidence that experimental inference of cancer-associated variants can accelerate the translation of cancer genome sequencing data. As the first pilot study of this scale, this study also provides valuable empirical knowledge on how to perform large-scale functional characterization of diverse cancer alleles.

The first important lesson from this study is the importance of selecting many alleles of the same gene for meaningful analysis of expression profiling data. Without at least six alleles from the same gene, including the wild type, it is difficult to see the pattern of clustering among the alleles. Including at least two biological replicates will greatly enhance the confidence in the

gene expression signature; two biological replicates should be required for the wild type allele at the least. In the study of 500 alleles, for example, it would be much more valuable to study 10 alleles of 50 genes than to study two alleles of 250 genes.

In line with the first point, it is critically important to include as many well-characterized alleles as possible, for these alleles will guide the interpretation of the other, unknown alleles. Moreover, it can be very valuable to study every single possible mutated allele in genes that are currently used in clinical setting to generate a complete dictionary. As these genes are well characterized, tailored phenotypic assays would be more appropriate than *in vivo* tumor formation or gene expression profiling.

The second lesson is the importance of testing the cellular background in eliciting the transcriptional impact of specific alleles. For example, the gene expression change induced by tumor suppressor genes such as *KEAP1* may only be distinctive enough to be detectable in *KEAP1* null background (Alice Berger, personal communication). In cells with endogenous wild type *KEAP1*, the difference between the impacts of overexpressed loss-of-function variant and wild type may be masked by the presence of the endogenous wild type allele. In light of this observation, the same set of alleles should be investigated in multiple cellular backgrounds. If a consensus gene expression signature can be derived from multiple backgrounds, it will provide a more powerful and accurate picture of specific variants.

The third lesson is an apparent but often-neglected point in large-scale ORF screens: the quality of ORFs. As we are attempting to study the changes associated with a single amino acid substitution, having concurrent alterations in the ORF would confound results. Constructing an ORF collection with 100% sequence accuracy is tremendously difficult as many genes have multiple transcripts and SNP variants. However, nearly all significantly mutated cancer genes have one reference sequence that the research community has reached a consensus on, which should be prioritized for study. In addition, large genes are typically vastly underrepresented in ORF libraries. For longer genes, individualized cloning strategies are required. These genes

may be unfit for the large-scale characterization described here due to inefficiencies in creating virus carrying large vectors and need to be studied individually. ORF library with 100% correct sequences would be a tremendous asset to the variant phenotyping community, not only in cancer but also in other diseases.

The pooling strategy was also found to be essential in this study. Pooling alleles in a greater number of combinations and testing each pool in fewer mice would be clearly more informative than smaller number of combinations injected into higher number of mice. For example, 28 pools injected into two mice per pool would allow discovery of more alleles than 14 pools injected into four mice per pool.

From the pooled *in vivo* screen, the "jackpot effect," in which the strongest alleles overtaking the entire tumor, was evident. Although the strong tumorigenic phenotype of potently transforming allele provides convincing evidence that it may be a driver allele in human cancer, this approach may not be the most efficient use of resources due to high false negative rates for alleles screened in the same pool as the jackpot allele. The jackpot effect is a consequence of the stochasticity of tumor formation in *in vivo* environment. This stochastic effect has minimal impact on strong oncogenic alleles such as activating *KRAS* alleles because the probability that a cell with this strong pro-tumorigenic allele would generate a tumor is fairly high, such that when more than a couple hundred cells are injected, the tumor take rate is ~100% since the take rate = 1-(1-probability of forming tumor)^(number of cells). However, for less robust alleles with orders of magnitude lower "probability of forming tumor," the take rate can be variable. Since the mouse needs to be sacrificed when the tumor size reaches predetermined size, variable take rate results in inconsistent scoring in a specific allele. To address this issue, I advocate using more democratic *in vitro* assays such as growing cells on low attachment plate to measure transformation phenotype.

Another important observation from the *in vivo* screen was the identification of false positives alleles, such as *FAM200A*$^{S481N}$ and *AKT1*$^{Q79K}$, which scored in pooled screen but did

not drive tumor formation when assessed individually. This was a puzzling observation as the number of cells with specific alleles injected in individual assay is orders of magnitude higher than the number of cells in the pooled assay. My conjecture is that, for cells overexpressing these false positive alleles, co-injection with cells transduced with different alleles could have facilitated tumorigenesis by paracrine effect of secretory factors from the neighboring cells. This hypothesis needs to be further validated, but the pool composition should be constructed with this effect in mind. The lessons described above will help improve sensitivity and specificity of studies of this kind in the future.

Based on the lessons I learned from this project, I propose a way forward to systematically characterize non-synonymous point mutations in cancer.

1. For clinically sequenced genes, whether the purpose is counseling or patient stratification for therapy or clinical trial enrollment, saturation mutagenesis and allele-by-allele functional characterization will prove immensely useful. These experiments are increasingly feasible as efficient and cost-effective construction of libraries with every possible amino acid substitution becomes more widely available. For small genes like *KRAS*, such libraries cost about $10,000 to synthesize. For larger genes like *EGFR*, the cost is roughly $40,000. Moreover, as the cancer-relevant functions of these genes are well characterized, the library can be screened using specific cancer-relevant functional assays.

2. For significantly mutated genes in cancer, about 200 genes, template ORFs with 100% sequence fidelity should be generated, and six to ten mutant alleles should be generated. Well-characterized and hot spot mutations should be prioritized and biological replicates should be included. These alleles should to be assessed using L1000 gene expression assays in multiple cell lines from different lineages. By analyzing the expression signature among the biological replicates, the cell line in which each gene is “readable” (provides meaningful expression profiles) should be identified. Based on the gene expression signature in the readable cell line, likely functional alleles can be differentiated from the likely neutral alleles.

3-1. For significantly mutated genes with likely functional alleles determined as above, all alleles created for this gene in 2. should to be subjected to *in vitro* transformation assays that measure anchorage independence.

3-2. For significantly mutated genes without likely functional alleles determined as above, gene expression may not be a suitable way to detect their functional impact. This category likely includes genes such as *POT1* and splicing factors. Genes in this category are likely not suitable for high throughput characterization and will need to be studied individually.

4. These characterization efforts need to be accompanied by development of better *in silico* methods. Methods described in 1. alone would likely generate enough training data to enable improvement in machine learning based variant calling algorithms such as Polyphen2. This will in turn teach us how to select features better and how to weigh features such as evolutionary conservation, biochemical properties of amino acid change and 3D spatial relationship correctly. One can speculate that after multiple iterations, the accuracy of *in silico* methods would be high enough that we may be able to reliably utilize *in silico* methods instead of experimental methods.

After 1-4, we will have a list of genes and alleles that we have enough confidence to embark on in-depth characterization that will eventually elucidate the full mechanism.

In summary, studies presented here describe approaches to translate structural cancer genome data into functional understanding. As future iterations of similar studies accumulate more functional data, we will be able to accurately describe and predict the functional consequences of novel alleles, systematically decreasing the number of variants of unknown significance.

# Appendix

The following tables are provided separately as Excel spreadsheets:

Supplementary Table S1 Genes and alleles selected for the project

Supplementary Table S2 Annotation of 1163 ORFs

Supplementary Table S3 Pool composition of in vivo screen

Supplementary Table S4 Composition of cells and tumors from the in vivo screen

Supplementary Table S5 L1000 gene expression data of 1036 ORFs

## Supplementary Table S6. Comparison to in silico methods.

| Mutation | This Study | Concordance to Polyphen2 | Concordance to Mutation Assessor | Concordance to CHASM | Concordance to VEST | Polyphen2 call | Mutation Assessor call | CHASM FDR (red<0.05) | VEST FDR (red<0.05) |
|---|---|---|---|---|---|---|---|---|---|
| AKT1_p.D44N | neutral | 0 | 1 | 1 | 1 | possibly damaging | neutral | >0.05 | >0.05 |
| AKT1_p.E17K | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| AKT1_p.E267G | functional | 0 | 0 | 1 | 1 | benign | neutral | <0.05 | <0.05 |
| AKT1_p.L52R | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| AKT1_p.Q79K | functional | 1 | 0 | 0 | 1 | probably damaging | low | >0.05 | <0.05 |
| AKT1_p.R370C | functional | 0 | 1 | 1 | 1 | benign | medium | <0.05 | <0.05 |
| AKT1_p.V201I | neutral | 1 | 1 | 0 | 1 | benign | neutral | <0.05 | >0.05 |
| FBXW7_p.A502V | neutral | 0 | 0 | 0 | 0 | probably damaging | medium | <0.05 | <0.05 |
| FBXW7_p.I347M | neutral | 1 | 1 | 1 | 0 | benign | neutral | >0.05 | <0.05 |
| FBXW7_p.R465C | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| FBXW7_p.R465H | functional | 1 | 0 | 1 | 1 | probably damaging | low | <0.05 | <0.05 |
| FBXW7_p.R505C | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| FBXW7_p.R658Q | neutral | 0 | 0 | 0 | 1 | probably damaging | medium | <0.05 | >0.05 |
| FBXW7_p.R689Q | neutral | 0 | 0 | 0 | 1 | probably damaging | medium | <0.05 | >0.05 |
| FBXW7_p.S462Y | neutral | 0 | 0 | 1 | 0 | probably damaging | medium | >0.05 | <0.05 |
| IDH1_p.E190K | neutral | 1 | 1 | 1 | 1 | benign | neutral | >0.05 | >0.05 |
| IDH1_p.P33S | neutral | 0 | 0 | 0 | 1 | probably damaging | medium | <0.05 | >0.05 |
| IDH1_p.R132C | functional | 0 | 1 | 1 | 1 | benign | high | <0.05 | <0.05 |
| IDH1_p.R132H | functional | 0 | 1 | 0 | 1 | benign | high | >0.05 | <0.05 |
| IDH1_p.R132L | functional | 0 | 1 | 1 | 1 | benign | high | <0.05 | <0.05 |
| IDH1_p.R132S | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| IDH2_p.A416V | neutral | 0 | 0 | 1 | 0 | probably damaging | medium | >0.05 | <0.05 |
| IDH2_p.A47V | neutral | 1 | 1 | 1 | 1 | benign | low | >0.05 | >0.05 |
| IDH2_p.E268D | neutral | 1 | 1 | 1 | 1 | benign | low | >0.05 | >0.05 |
| IDH2_p.G137E | neutral | 0 | 0 | 1 | 0 | probably damaging | high | >0.05 | <0.05 |
| IDH2_p.I139F | neutral | 0 | 0 | 0 | 0 | possibly damaging | high | <0.05 | <0.05 |
| IDH2_p.R172K | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| IDH2_p.R172M | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| IDH2_p.T331M | neutral | 0 | 0 | 0 | 0 | probably damaging | high | <0.05 | <0.05 |
| KRAS_p.A59G | functional | 1 | 1 | 0 | 1 | possibly damaging | high | >0.05 | <0.05 |
| KRAS_p.D33E | functional | 1 | 0 | 1 | 1 | probably damaging | low | <0.05 | <0.05 |
| KRAS_p.E62K | functional | 1 | 1 | 0 | 1 | possibly damaging | high | >0.05 | <0.05 |
| KRAS_p.G12A | functional | 1 | 1 | 1 | 1 | possibly damaging | medium | <0.05 | <0.05 |
| KRAS_p.G12C | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| KRAS_p.G12D | functional | 1 | 1 | 1 | 1 | possibly damaging | medium | <0.05 | <0.05 |
| KRAS_p.G12S | functional | 1 | 0 | 1 | 1 | possibly damaging | low | <0.05 | <0.05 |
| KRAS_p.G12V | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| NFE2L2_p.G31A | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| NFE2L2_p.G31R | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| NFE2L2_p.G31V | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| NFE2L2_p.N160S | neutral | 1 | 1 | 1 | 1 | benign | neutral | >0.05 | >0.05 |
| NFE2L2_p.T80K | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| NRAS_p.G12A | functional | 1 | 1 | 1 | 1 | possibly damaging | medium | <0.05 | <0.05 |
| NRAS_p.G12C | functional | 1 | 1 | 1 | 1 | possibly damaging | medium | <0.05 | <0.05 |
| NRAS_p.Q61H | functional | 0 | 1 | 1 | 1 | benign | high | <0.05 | <0.05 |
| NRAS_p.Q61K | functional | 1 | 1 | 1 | 1 | possibly damaging | high | <0.05 | <0.05 |
| NRAS_p.Q61L | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| NRAS_p.Q61R | functional | 0 | 1 | 1 | 1 | benign | medium | <0.05 | <0.05 |
| NRAS_p.Y64D | neutral | 0 | 0 | 1 | 0 | probably damaging | medium | >0.05 | <0.05 |
| PIK3CB_p.A1048V | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| PIK3CB_p.A593V | neutral | 1 | 1 | 0 | 0 | benign | low | <0.05 | <0.05 |
| PIK3CB_p.E497D | neutral | 1 | 1 | 1 | 1 | benign | neutral | >0.05 | >0.05 |
| POT1_p.G76V | functional | 1 | 1 | 0 | 1 | probably damaging | medium | >0.05 | <0.05 |
| POT1_p.L265H | neutral | 0 | 0 | 1 | 0 | probably damaging | medium | >0.05 | <0.05 |
| POT1_p.L388M | neutral | 1 | 1 | 1 | 1 | benign | low | >0.05 | >0.05 |
| PTEN_p.F90S | neutral | 0 | 0 | 0 | 0 | probably damaging | high | <0.05 | <0.05 |
| PTEN_p.G127R | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| PTEN_p.G127V | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| PTEN_p.G129E | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| PTEN_p.G129V | functional | 1 | 1 | 1 | 1 | probably damaging | high | <0.05 | <0.05 |
| PTEN_p.K6N | neutral | 1 | 0 | 0 | 1 | benign | medium | <0.05 | >0.05 |
| PTEN_p.R173H | neutral | 0 | 0 | 0 | 0 | probably damaging | medium | <0.05 | <0.05 |
| PTEN_p.R233Q | neutral | 0 | 0 | 0 | 0 | possibly damaging | medium | <0.05 | <0.05 |
| SPOP_p.E47A | functional | 1 | 0 | 0 | 1 | probably damaging | low | >0.05 | <0.05 |
| SPOP_p.E50K | functional | 0 | 0 | 1 | 0 | benign | low | <0.05 | <0.05 |
| SPOP_p.F102C | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| SPOP_p.F133S | functional | 1 | 1 | 1 | 1 | probably damaging | medium | <0.05 | <0.05 |
| SPOP_p.K101I | neutral | 0 | 0 | 1 | 0 | probably damaging | medium | >0.05 | <0.05 |
| SPOP_p.K134N | functional | 1 | 0 | 1 | 0 | probably damaging | low | <0.05 | >0.05 |
| SPOP_p.W131C | functional | 1 | 1 | 0 | 1 | probably damaging | medium | >0.05 | <0.05 |
| SPOP_p.W131G | funcitonal | 1 | 1 | 1 | 1 | possibly damaging | high | <0.05 | <0.05 |
| | **Sum** | 47 | 47 | 52 | 55 | | | | |
| | **Percent Co** | 66.1971831 | 66.1971831 | 73.23943662 | 77.46478873 | | | | |

## Supplementary Table S7. Genes in 12q15 Amplified region.

| Ovarian | | chr12:69692322-71120515 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| #bin | name | chrom | strand | txStart | txEnd | cdsStart | cdsEnd | exonCount | exonStarts | exonEnds | name2 |
| 1117 | ENST00000261267.2 | chr12 | + | 69742120 | 69748014 | 69742188 | 69746999 | 4 | 69742120,697 | 69742324,697 | LYZ |
| 1117 | ENST00000549690.1 | chr12 | + | 69742163 | 69747275 | 69742188 | 69747045 | 3 | 69742163,697 | 69742324,697 | LYZ |
| 1117 | ENST00000548839.1 | chr12 | + | 69742165 | 69744286 | 69742188 | 69744066 | 2 | 69742165,697 | 69742324,697 | LYZ |
| 1117 | ENST00000548900.1 | chr12 | - | 69747272 | 69748005 | 69748005 | 69748005 | 2 | 69747272,697 | 69747388,697 | RP11-1143G9.4 |
| 1117 | ENST00000247843.2 | chr12 | + | 69753482 | 69784576 | 69753752 | 69784096 | 7 | 69753482,697 | 69753803,697 | YEATS4 |
| 1117 | ENST00000548020.1 | chr12 | + | 69753489 | 69784411 | 69753752 | 69784096 | 5 | 69753489,697 | 69753803,697 | YEATS4 |
| 1117 | ENST00000549261.1 | chr12 | + | 69843249 | 69854504 | 69854504 | 69854504 | 3 | 69843249,698 | 69843468,698 | RP11-956E11.1 |
| 1118 | ENST00000299293.2 | chr12 | + | 69864128 | 69973559 | 69962810 | 69968735 | 10 | 69864128,699 | 69864310,699 | FRS2 |
| 1118 | ENST00000549921.1 | chr12 | + | 69864154 | 69968744 | 69962810 | 69968735 | 9 | 69864154,699 | 69864310,699 | FRS2 |
| 1118 | ENST00000550389.1 | chr12 | + | 69864185 | 69973562 | 69962810 | 69968735 | 7 | 69864185,699 | 69864310,699 | FRS2 |
| 1118 | ENST00000397997.2 | chr12 | + | 69924602 | 69973559 | 69962810 | 69968735 | 8 | 69924602,699 | 69924740,699 | FRS2 |
| 139 | ENST00000299300.6 | chr12 | + | 69979113 | 69995350 | 69979301 | 69995105 | 16 | 69979113,699 | 69979304,699 | CCT2 |
| 139 | ENST00000544368.2 | chr12 | + | 69979239 | 69995305 | 69979301 | 69995232 | 15 | 69979239,699 | 69979304,699 | CCT2 |
| 139 | ENST00000543146.2 | chr12 | + | 69979445 | 69995345 | 69980595 | 69995105 | 16 | 69979445,699 | 69979789,699 | CCT2 |
| 1118 | ENST00000550871.1 | chr12 | + | 69982764 | 69985840 | 69985840 | 69985840 | 3 | 69982764,699 | 69982851,699 | CCT2 |
| 1119 | ENST00000361484.3 | chr12 | - | 70002350 | 70004942 | 70003784 | 70004618 | 1 | 70002350, | 70004942, | LRRC10 |
| 1119 | ENST00000331471.4 | chr12 | - | 70037333 | 70093141 | 70037470 | 70091578 | 10 | 70037333,700 | 70037567,700 | BEST3 |
| 1119 | ENST00000488961.1 | chr12 | - | 70047388 | 70083056 | 70048686 | 70072668 | 6 | 70047388,700 | 70049593,700 | BEST3 |
| 1119 | ENST00000330891.5 | chr12 | - | 70047388 | 70093196 | 70048686 | 70091578 | 10 | 70047388,700 | 70049593,700 | BEST3 |
| 1119 | ENST00000553096.1 | chr12 | - | 70047598 | 70093065 | 70048686 | 70087616 | 8 | 70047598,700 | 70049593,700 | BEST3 |
| 1119 | ENST00000476098.1 | chr12 | - | 70063458 | 70093131 | 70064225 | 70072668 | 7 | 70063458,700 | 70064358,700 | BEST3 |
| 1119 | ENST00000266661.4 | chr12 | - | 70077018 | 70093256 | 70078332 | 70087616 | 3 | 70077018,700 | 70078388,700 | BEST3 |
| 1119 | ENST00000551160.1 | chr12 | - | 70078187 | 70093175 | 70078332 | 70087616 | 4 | 70078187,700 | 70078388,700 | BEST3 |
| 1119 | ENST00000393365.1 | chr12 | - | 70078332 | 70093141 | 70078332 | 70087616 | 5 | 70078332,700 | 70078388,700 | BEST3 |
| 1119 | ENST00000533674.1 | chr12 | - | 70081038 | 70093124 | 70093124 | 70093124 | 4 | 70081038,700 | 70081239,700 | BEST3 |
| 139 | ENST00000501387.1 | chr12 | - | 70107412 | 70132348 | 70132348 | 70132348 | 4 | 70107412,701 | 70109367,701 | RP11-588G21.2 |
| 139 | ENST00000501300.1 | chr12 | - | 70116101 | 70132342 | 70132342 | 70132342 | 3 | 70116101,701 | 70116351,701 | RP11-588G21.2 |
| 1120 | ENST00000247833.7 | chr12 | + | 70132460 | 70211157 | 70149188 | 70209226 | 11 | 70132460,701 | 70132811,701 | RAB3IP |
| 1120 | ENST00000378815.6 | chr12 | + | 70132641 | 70190427 | 70149188 | 70189184 | 6 | 70132641,701 | 70132811,701 | RAB3IP |
| 1120 | ENST00000483530.2 | chr12 | + | 70132641 | 70209503 | 70149188 | 70206801 | 10 | 70132641,701 | 70132811,701 | RAB3IP |
| 1120 | ENST00000325555.9 | chr12 | + | 70132641 | 70209503 | 70188228 | 70209226 | 11 | 70132641,701 | 70132811,701 | RAB3IP |
| 1120 | ENST00000550536.1 | chr12 | + | 70133169 | 70216984 | 70133626 | 70209226 | 11 | 70133169,701 | 70133649,701 | RAB3IP |
| 1120 | ENST00000362025.5 | chr12 | + | 70133179 | 70209503 | 70133626 | 70206801 | 10 | 70133179,701 | 70133649,701 | RAB3IP |
| 1120 | ENST00000551641.1 | chr12 | + | 70172729 | 70210961 | 70188228 | 70209226 | 9 | 70172729,701 | 70172961,701 | RAB3IP |
| 1120 | ENST00000553099.1 | chr12 | + | 70172746 | 70210897 | 70188228 | 70209226 | 9 | 70172746,701 | 70172876,701 | RAB3IP |
| 1120 | ENST00000550847.1 | chr12 | + | 70190352 | 70209330 | 70190414 | 70209226 | 6 | 70190352,701 | 70190423,701 | RAB3IP |
| 1120 | ENST00000550437.1 | chr12 | + | 70195448 | 70249143 | 70195448 | 70219110 | 5 | 70195448,702 | 70195501,702 | AC025263.3 |
| 17 | ENST00000552032.1 | chr12 | + | 70219083 | 70352503 | 70284895 | 70352311 | 25 | 70219083,702 | 70219343,702 | C12orf28 |
| 1121 | ENST00000299350.4 | chr12 | + | 70320436 | 70352503 | 70326367 | 70352311 | 12 | 70320436,703 | 70320514,703 | C12orf28 |
| 1121 | ENST00000535034.1 | chr12 | + | 70326315 | 70352387 | 70326367 | 70352311 | 9 | 70326315,703 | 70326378,703 | C12orf28 |
| 1121 | ENST00000547547.1 | chr12 | - | 70340322 | 70340861 | 70340861 | 70340861 | 2 | 70340322,703 | 70340575,703 | RP11-611E13.3 |
| 1123 | ENST00000552324.1 | chr12 | + | 70574117 | 70595784 | 70595784 | 70595784 | 3 | 70574117,705 | 70574318,705 | RP11-320P7.2 |
| 1123 | ENST00000552998.1 | chr12 | - | 70612911 | 70615642 | 70615642 | 70615642 | 2 | 70612911,706 | 70613377,706 | RP11-320P7.1 |
| 1123 | ENST00000549651.1 | chr12 | - | 70636085 | 70637140 | 70637140 | 70637140 | 2 | 70636085,706 | 70636673,706 | RP11-611E13.2 |
| 140 | ENST00000229195.3 | chr12 | + | 70636776 | 70748773 | 70672006 | 70747695 | 16 | 70636776,706 | 70637260,706 | CNOT2 |
| 140 | ENST00000418359.3 | chr12 | + | 70636809 | 70748773 | 70672006 | 70747695 | 17 | 70636809,706 | 70637017,706 | CNOT2 |
| 1124 | ENST00000548230.1 | chr12 | + | 70721286 | 70729246 | 70729246 | 70729246 | 5 | 70721286,707 | 70721495,707 | CNOT2 |
| 1124 | ENST00000551483.1 | chr12 | + | 70728214 | 70747717 | 70732471 | 70747695 | 7 | 70728214,707 | 70732343,707 | CNOT2 |
| 140 | ENST00000258111.4 | chr12 | + | 70760055 | 70828072 | 70760514 | 70824433 | 3 | 70760055,707 | 70760850,707 | KCNMB4 |
| 1125 | ENST00000410473.1 | chr12 | - | 70837563 | 70837703 | 70837703 | 70837703 | 1 | 70837563, | 70837703, | U4 |
| 140 | ENST00000549460.1 | chr12 | + | 70861859 | 70931840 | 70931840 | 70931840 | 6 | 70861859,708 | 70862107,708 | RP11-588H23.3 |
| 140 | ENST00000548687.1 | chr12 | + | 70861864 | 70932859 | 70932859 | 70932859 | 9 | 70861864,709 | 70862107,709 | RP11-588H23.3 |
| 1125 | ENST00000548924.1 | chr12 | + | 70861889 | 70905002 | 70905002 | 70905002 | 5 | 70861889,709 | 70862107,709 | RP11-588H23.3 |
| 140 | ENST00000549616.1 | chr12 | + | 70861902 | 70914619 | 70914619 | 70914619 | 6 | 70861902,709 | 70862107,709 | RP11-588H23.3 |
| 140 | ENST00000549359.1 | chr12 | + | 70861902 | 70921529 | 70921529 | 70921529 | 6 | 70861902,709 | 70862107,709 | RP11-588H23.3 |
| 140 | ENST00000551438.1 | chr12 | + | 70861974 | 70931985 | 70931985 | 70931985 | 6 | 70861974,709 | 70862107,709 | RP11-588H23.3 |
| 1126 | ENST00000451516.2 | chr12 | - | 70910629 | 71003594 | 70915268 | 71003594 | 31 | 70910629,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000334414.6 | chr12 | - | 70910629 | 71031220 | 70915268 | 71031175 | 34 | 70910629,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000547656.1 | chr12 | + | 70913970 | 70932279 | 70932279 | 70932279 | 2 | 70913970,709 | 70914047,709 | RP11-588H23.3 |
| 1126 | ENST00000546836.1 | chr12 | + | 70913985 | 70932443 | 70932443 | 70932443 | 3 | 70913985,709 | 70914047,709 | RP11-588H23.3 |
| 1126 | ENST00000550358.1 | chr12 | - | 70915095 | 71031201 | 70915268 | 71031175 | 33 | 70915095,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000544694.1 | chr12 | - | 70915096 | 71031201 | 70965684 | 71031175 | 34 | 70915096,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000538708.1 | chr12 | - | 70915182 | 71003623 | 70915268 | 71003594 | 31 | 70915182,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000550857.1 | chr12 | - | 70915182 | 71003623 | 70915268 | 71003594 | 31 | 70915182,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000261266.5 | chr12 | - | 70915182 | 71003624 | 70915268 | 71003594 | 32 | 70915182,709 | 70915291,709 | PTPRB |
| 1126 | ENST00000551525.1 | chr12 | - | 70952567 | 71031200 | 70953081 | 71031175 | 18 | 70952567,709 | 70953404,709 | PTPRB |
| 1126 | ENST00000538174.2 | chr12 | - | 70978744 | 71031194 | 71031194 | 71031194 | 10 | 70978744,709 | 70979075,709 | PTPRB |
| 140 | ENST00000440835.2 | chr12 | - | 71031852 | 71148441 | 71032963 | 71147973 | 10 | 71031852,710 | 71033057,710 | PTPRR |
| 140 | ENST00000537619.2 | chr12 | - | 71031858 | 71058457 | 71058457 | 71058457 | 5 | 71031858,710 | 71033057,710 | PTPRR |
| 140 | ENST00000378778.1 | chr12 | - | 71031861 | 71148373 | 71032963 | 71148373 | 11 | 71031861,710 | 71033057,710 | PTPRR |
| 17 | ENST00000283228.2 | chr12 | - | 71031861 | 71314623 | 71032963 | 71314170 | 14 | 71031861,710 | 71033057,710 | PTPRR |
| 140 | ENST00000342084.4 | chr12 | - | 71032710 | 71182762 | 71032963 | 71182616 | 13 | 71032710,710 | 71033057,710 | PTPRR |
| 140 | ENST00000549308.1 | chr12 | - | 71032839 | 71148496 | 71032963 | 71147973 | 11 | 71032839,710 | 71033057,710 | PTPRR |

**Supplementary Table Legend**

**Supplementary Table S1: Genes and alleles selected for the project.**
This table includes description of all the alleles selected for the project, including the ones excluded due to template unavailability and unsuccessful mutagenesis. The meaning of column headings is specified below:

- template_available: TRUE, when the template ORF was available in hORFeme 8.1 collection
- mutagenesis_successful: TRUE, when the mutagenesis was successful
- BarcodedVectorID: identification number given to each vector. (failed_QC: sequencing results were not satisfactory, template_unavailable: template was not available)
- n_AML - n_UCEC: number of times each mutation was found in each cancer type
- n_pancan: sum of columns of n_AML - n_UCEC. This column was used for generating Fig. 1B.

**Supplementary Table S2: Annotation of 1163 ORFs.**
This table includes description of all the alleles used in the *in vivo* screening and gene expression experiments. Only the mutant alleles (PC_MUT, under category) were included in the in vivo screening (474 total alleles). All of the ORFs were included for the gene expression assay. The meaning of column headings are specified below:

- plate_well_ID: identification number given to each well of the assay plate. This ID is used in Supplementary Table S5.
- clone_ID: identification number identical to BarcodedVectorID in Supplementary Table S1.
- Vector: lentiviral vectors used. PLX_TRC317 is identical to pLEX_307 (https://www.addgene.org/41392/). It has EF1α promoter and puromycin selection marker. PLX_TRC304 is identical to pLX304 (https://www.addgene.org/25890/). It has CMV promoter and blasticidin selection marker.
- open_close: when the C-terminal of the ORFs did not have the stop codon, it resulted in V5 tagging at the C-terminal (annotated as “open”). “close” otherwise.
- gene, protein_change: shows gene and protein change.
- point_mutation: additional point mutation found. “c.262C>T|p.H88Y” shows that nucleotide position 262 was T, not C, which resulted in non-synonymous mutation H88Y.
- indel: additional insertion or deletion found. “1121delG” means nucleotide position 1121 had a single G deletion.
- intended_transcript: shows the intended RefSeq accession number.
- category:
    - PC_MUT: mutant alleles generated for the study. 474 in total.
    - PC_WT: wild type alleles generated for the study. 187 unique alleles, 334 in total due to many alleles having two entries (open and close forms).
    - REF: reference alleles of known biological function. 232 unique alleles, 308 in total due to many alleles having more than one entry.
    - CTL_INRT: negative controls including BFP, eGFP, HcRED, LacZ, and Luciferase. 5 unique alleles, 35 in total due to each allele being included seven times.
    - CTL_L1000: internal expression control for L1000 assay, including DNMT3A, NFE2L2, NFKBIA, RHEB. 4 unique alleles, 12 total due to each allele being included three times.
    - infection_efficiency: infection efficiency shown in percentage. Refer to Methods for full details.

**Supplementary Table S3: Pool composition of *in vivo* screen.**
This table shows the composition 14 pools. The first column shows the name of the mutant or control alleles. TRUE mean that the allele belongs to that pool. For example, "A4GALT_p.A272V" belongs to Pool 5 and Pool 14. To search for alleles in each pool, use the filtering function of the Excel (shown as funnel shaped icon).

**Supplementary Table S4: Composition of cells and tumors from the *in vivo* screen.**
These tables show the composition of pre-expansion and pre-injection cells and tumors in each pool. The numbers are shown in percentage.

- Supplementary Table S4-1: Composition of Pre-expansion cell culture. Supplementary Table S3 describes the pool membership of each Mutation (first column). This table shows the barcode representation immediately after pooling the cells after arrayed infection.
- Supplementary Table S4-2: Composition of Pre-injection cell culture. Supplementary Table S3 describes the pool membership of each Mutation (first column). This table shows the barcode representation after 15 days of *in vitro* culture, right before cells were injected into nude mice. All enrichment analysis was done using this as reference point.
- Supplementary Table S4-3 - Supplementary Table S4-14: Composition of each tumor in the Pool1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13 in that order. Second column to last column headings show tumor ID. Tumor ID "P1M1_L" means Pool1, mouse 1, left flank injection site. "L": left flank, "R": right flank, "T": upper back.

**Supplementary Table S5: L1000 gene expression data of 1036 ORFs.**
This table shows the L1000 gene expression date of 1036 ORFs that passed 40% infection efficiency cutoff.

- landmark: this column shows the 978 landmark genes, whose expressions are measured in L1000 assay.
- second column to last column: these columns show the plate_well_ID, as specified in Supplementary Table S2.

**Supplementary Table S6: Comparison to *in silico* methods.**
This table shows the calls of four different in silico methods (Polyphen2, Mutation Assessor, CHASM, and VEST) and comparison to our results. See the methods for description.

- Mutation: lists alleles
- This Study: functional description from this study. "functional" denotes both gain-of-function and loss-of-function alleles. "neutral" denotes likely passenger mutations.
- Concordance to Polyphen2, Mutation Assessor, CHASM, VEST: "1" if concordant, "0" otherwise.
- Polyphen2 score, Polyphen2 call: output from Polyphen2.
- Mutation Assessor score, Mutation Assessor call: output from Mutation Assessor
- CHASM cancer driver p-value (missense), CHASM FDR (red<0.05): output from CHASM. FDR <0.05 was colored red.
- VEST pathogenicity p-value (non-silent), VEST FDR (red<0.05): output from VEST. FDR <0.05 was colored red.

# References

1. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med 2011;17(3):297-303.

2. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature Reviews Cancer 2004;4(3):177-83.

3. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458(7239):719-24.

4. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst 1960;25:85-109.

5. Rowley JD. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243(5405):290-3.

6. King CR, Kraus MH, Aaronson SA. Amplification of a novel v-erbB-related gene in a human mammary carcinoma. Science 1985;229(4717):974-6.

7. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature 2002;417(6892):949-54.

8. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 2001;344(14):1031-7.

9. Flaherty KT, Puzanov I, Kim KB, Ribas A, McArthur GA, Sosman JA, et al. Inhibition of Mutated, Activated BRAF in Metastatic Melanoma. New England Journal of Medicine 2010;363(9):809-19.

10. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 2001;344(11):783-92.

11. Martin GS. The road to Src. Oncogene 2004;23(48):7910-7.

12. Cox AD, Der CJ. Ras history: The saga continues. Small GTPases 2010;1(1):2-27.

13. Slamon DJ. Proto-oncogenes and human cancers. N Engl J Med 1987;317(15):955-7.

14. Tabin CJ, Bradley SM, Bargmann CI, Weinberg RA, Papageorge AG, Scolnick EM, et al. Mechanism of activation of a human oncogene. Nature 1982;300(5888):143-9.

15. Reddy EP, Reynolds RK, Santos E, Barbacid M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature 1982;300(5888):149-52.

16. Schwab M, Alitalo K, Klempnauer KH, Varmus HE, Bishop JM, Gilbert F, et al. Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature 1983;305(5931):245-8.

17. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007;7(4):233-45.

18. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310(5748):644-48.

19. Westbrook TF, Martin ES, Schlabach MR, Leng YM, Liang AC, Feng B, et al. A genetic screen for candidate tumor suppressors identifies REST. Cell 2005;121(6):837-48.

20. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448(7153):561-6.

21. Boehm JS, Zhao JJ, Yao J, Kim SY, Firestein R, Dunn IF, et al. Integrative genomic approaches identify IKBKE as a breast cancer oncogene. Cell 2007;129(6):1065-79.

22. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.

23. Stephens P, Hunter C, Bignell G, Edkins S, Davies H, Teague J, et al. Intragenic ERBB2 kinase mutations in tumours. Nature 2004;431(7008):525-26.

24. Bachman KE, Argani P, Samuels Y, Silliman N, Ptak J, Szabo S, et al. The PIK3CA gene is mutated with high frequency in human breast cancers. Cancer Biology & Therapy 2004;3(8):772-75.

25. Samuels Y, Wang ZH, Bardelli A, Silliman N, Ptak J, Szabo S, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science 2004;304(5670):554-54.

26. Levine RL, Wadleigh M, Cools J, Ebert BL, Wernig G, Huntly BJP, et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell 2005;7(4):387-97.

27. Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, et al. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 2007;448(7152):439-44.

28. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007;318(5853):1108-13.

29. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321(5897):1807-12.

30. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26(10):1135-45.

31. Garraway LA, Lander ES. Lessons from the Cancer Genome. Cell 2013;153(1):17-37.

32. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science 2013;339(6127):1546-58.

33. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, et al. International network of cancer genome projects. Nature 2010;464(7291):993-8.

34. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. Nat Rev Genet 2016;17(2):93-108.

35. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell 2013;153(3):666-77.

36. Stephens PJ, Greenman CD, Fu BY, Yang FT, Bignell GR, Mudie LJ, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell 2011;144(1):27-40.

37. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer 2012;12(10):663-70.

38. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. Science 2013;339(6122):957-9.

39. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, et al. TERT Promoter Mutations in Familial and Sporadic Melanoma. Science 2013;339(6122):959-61.

40. Borah S, Xi L, Zaug AJ, Powell NM, Dancik GM, Cohen SB, et al. Cancer. TERT promoter mutations and telomerase reactivation in urothelial cancer. Science 2015;347(6225):1006-10.

41. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet 2014;46(11):1160-5.

42. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 2011;12(10):683-91.

43. Gartner JJ, Parker SC, Prickett TD, Dutton-Regester K, Stitzel ML, Lin JC, et al. Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. Proc Natl Acad Sci U S A 2013;110(33):13481-6.

44. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell 2014;156(6):1324-35.

45. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43(Database issue):D805-11.

46. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science 2015;349(6255):1483-9.

47. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013.

48. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, et al. A landscape of driver mutations in melanoma. Cell 2012;150(2):251-63.

49. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res 2012;22(8):1589-98.

50. Pon JR, Marra MA. Driver and passenger mutations in cancer. Annu Rev Pathol 2015;10:25-50.

51. Getz G, Hofling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, et al. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 2007;317(5844):1500.

52. Rubin AF, Green P. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 2007;317(5844):1500.

53. Forrest WF, Cavet G. Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 2007;317(5844):1500; author reply 00.

54. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med 2014;6(1):5.

55. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 2011;12(9):628-40.

56. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7(4):248-9.

57. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39(17):e118.

58. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009;69(16):6660-7.

59. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 2013;14 Suppl 3:S3.

60. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 2012;40(Web Server issue):W452-7.

61. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, et al. Computational approaches to identify functional genetic variants in cancer genomes. Nat Methods 2013;10(8):723-9.

62. Hocking CM, Price TJ. Panitumumab in the management of patients with KRAS wild-type metastatic colorectal cancer. Therap Adv Gastroenterol 2014;7(1):20-37.

63. Peeters M, Douillard JY, Van Cutsem E, Siena S, Zhang K, Williams R, et al. Mutant KRAS codon 12 and 13 alleles in patients with metastatic colorectal cancer: assessment as prognostic and predictive biomarkers of response to panitumumab. J Clin Oncol 2013;31(6):759-65.

64. Janku F, Hong DS, Fu S, Piha-Paul SA, Naing A, Falchook GS, et al. Assessing PIK3CA and PTEN in early-phase trials with PI3K/AKT/mTOR inhibitors. Cell Rep 2014;6(2):377-87.

65. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013;45(10):1134-40.

66. Collins S, Groudine M. Amplification of endogenous myc-related DNA sequences in a human myeloid leukaemia cell line. Nature 1982;298(5875):679-81.

67. Little CD, Nau MM, Carney DN, Gazdar AF, Minna JD. Amplification and expression of the c-myc oncogene in human lung cancer cell lines. Nature 1983;306(5939):194-6.

68. Hynes NE, Lane HA. ERBB receptors and cancer: The complexity of targeted inhibitors. Nature Reviews Cancer 2005;5(5):341-54.

69. Santarius T, Shipley J, Brewer D, Stratton M, Cooper C. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 2010;10:59 - 64.

70. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12(4):R41.

71. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A 2007;104(50):20007-12.

72. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463(7283):899-905.

73. Garraway LA, Widlund HR, Rubin MA, Getz G, Berger AJ, Ramaswamy S, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 2005;436(7047):117-22.

74. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet 2009;41(11):1238-42.

75. Theurillat JP, Metzler SC, Henzi N, Djouder N, Helbling M, Zimmermann AK, et al. URI is an oncogene amplified in ovarian cancer cells and is required for their survival. Cancer Cell 2011;19(3):317-32.

76. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci U S A 2008;105(51):20380-5.

77. Cheung HW, Du J, Boehm JS, He F, Weir BA, Wang X, et al. Amplification of CRKL induces transformation and epidermal growth factor receptor inhibitor resistance in human non-small cell lung cancers. Cancer Discov 2011;1(7):608-25.

78. Kim YH, Kwei KA, Girard L, Salari K, Kao J, Pacyna-Gengelbach M, et al. Genomic and functional analysis identifies CRKL as an oncogene amplified in lung cancer. Oncogene 2010;29(10):1421-30.

79. Sawey ET, Chanrion M, Cai C, Wu G, Zhang J, Zender L, et al. Identification of a therapeutic strategy targeting amplified FGF19 in liver cancer by Oncogenomic screening. Cancer Cell 2011;19(3):347-58.

80. Hagerstrand D, Tong A, Schumacher SE, Ilic N, Shen RR, Cheung HW, et al. Systematic interrogation of 3q26 identifies TLOC1 and SKIL as cancer drivers. Cancer Discov 2013.

81. Neupane M, Clark AP, Landini S, Birkbak NJ, Eklund AC, Lim E, et al. MECP2 Is a Frequently Amplified Oncogene with a Novel Epigenetic Mechanism That Mimics the Role of Activated RAS in Malignancy. Cancer Discov 2016;6(1):45-58.

82. Papa A, Wan L, Bonora M, Salmena L, Song MS, Hobbs RM, et al. Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function. Cell 2014;157(3):595-610.

83. Han SY, Kato H, Kato S, Suzuki T, Shibata H, Ishii S, et al. Functional evaluation of PTEN missense mutations using in vitro phosphoinositide phosphatase assay. Cancer Res 2000;60(12):3147-51.

84. Myers MP, Pass I, Batty IH, Van der Kaay J, Stolarov JP, Hemmings BA, et al. The lipid phosphatase activity of PTEN is critical for its tumor supressor function. Proc Natl Acad Sci U S A 1998;95(23):13513-8.

85. Nguyen HN, Yang JM, Jr., Rahdar M, Keniry M, Swaney KF, Parsons R, et al. A new class of cancer-associated PTEN mutations defined by membrane translocation defects. Oncogene 2015;34(28):3737-43.

86. Gremer L, Merbitz-Zahradnik T, Dvorsky R, Cirstea IC, Kratz CP, Zenker M, et al. Germline KRAS Mutations Cause Aberrant Biochemical and Physical Properties Leading to Developmental Disorders. Hum Mutat 2011;32(1):33-43.

87. Janakiraman M, Vakiani E, Zeng Z, Pratilas CA, Taylor BS, Chitale D, et al. Genomic and biological characterization of exon 4 KRAS mutations in human cancer. Cancer Res 2010;70(14):5901-11.

88. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646-74.

89. Mansour SL, Thomas KR, Capecchi MR. Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. Nature 1988;336(6197):348-52.

90. Frese KK, Tuveson DA. Maximizing mouse cancer models. Nat Rev Cancer 2007;7(9):645-58.

91. Jacks T, Shih TS, Schmitt EM, Bronson RT, Bernards A, Weinberg RA. Tumour predisposition in mice heterozygous for a targeted mutation in Nf1. Nat Genet 1994;7(3):353-61.

92. Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA, Jr., Butel JS, et al. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature 1992;356(6366):215-21.

93. Chin L, Tam A, Pomerantz J, Wong M, Holash J, Bardeesy N, et al. Essential role for oncogenic Ras in tumour maintenance. Nature 1999;400(6743):468-72.

94. Gossen M, Freundlieb S, Bender G, Muller G, Hillen W, Bujard H. Transcriptional activation by tetracyclines in mammalian cells. Science 1995;268(5218):1766-9.

95. Kistner A, Gossen M, Zimmermann F, Jerecic J, Ullmer C, Lubbert H, et al. Doxycycline-mediated quantitative and tissue-specific control of gene expression in transgenic mice. Proc Natl Acad Sci U S A 1996;93(20):10933-8.

96. Ventura A, Kirsch DG, McLaughlin ME, Tuveson DA, Grimm J, Lintault L, et al. Restoration of p53 function leads to tumour regression in vivo. Nature 2007;445(7128):661-65.

97. Sinn E, Muller W, Pattengale P, Tepler I, Wallace R, Leder P. Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo. Cell 1987;49(4):465-75.

98. Engelman JA, Chen L, Tan XH, Crosby K, Guimaraes AR, Upadhyay R, et al. Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. Nature Medicine 2008;14(12):1351-56.

99. Eccles DM, Mitchell G, Monteiro AN, Schmutzler R, Couch FJ, Spurdle AB, et al. BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. Ann Oncol 2015;26(10):2057-65.

100. Allegra CJ, Rumble RB, Hamilton SR, Mangu PB, Roach N, Hantel A, et al. Extended RAS Gene Mutation Testing in Metastatic Colorectal Carcinoma to Predict Response to Anti–Epidermal Growth Factor Receptor Monoclonal Antibody Therapy: American Society of Clinical Oncology Provisional Clinical Opinion Update 2015. Journal of Clinical Oncology 2015.

101. Moyer VA, Force USPST. Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer in Women: US Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine 2014;160(4):271-81.

102. Martin-Algarra S, Fernandez-Figueras MT, Lopez-Martin JA, Santos-Briz A, Arance A, Lozano MD, et al. Guidelines for biomarker testing in metastatic melanoma: a National Consensus of the Spanish Society of Pathology and the Spanish Society of Medical Oncology. Clin Transl Oncol 2014;16(4):362-73.

103. Lindeman NI, Cagle PT, Beasley MB, Chitale DA, Dacic S, Giaccone G, et al. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. J Thorac Oncol 2013;8(7):823-59.

104. Carlson R. The changing economics of DNA synthesis. Nat Biotech 2009;27(12):1091-94.

105. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res 2014;42(14):e112.

106. Yang X, Boehm JS, Yang X, Salehi-Ashtiani K, Hao T, Shen Y, et al. A public genome-scale lentiviral expression library of human ORFs. Nat Methods 2011;8(8):659-61.

107. Boehm JS, Hahn WC. Towards systematic functional characterization of cancer genomes. Nat Rev Genet 2011;12(7):487-98.

108. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 2009;462(7269):108-12.

109. Kaelin WG, Jr. Molecular biology. Use and abuse of RNAi to study mammalian gene function. Science 2012;337(6093):421-2.

110. Jackson AL, Burchard J, Schelter J, Chau BN, Cleary M, Lim L, et al. Widespread siRNA “off-target” transcript silencing mediated by seed region sequence complementarity. Rna 2006;12(7):1179-87.

111. Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res 2013;23(4):665-78.

112. Buehler E, Chen YC, Martin S. C911: A bench-level control for sequence specific siRNA off-target effects. PLoS One 2012;7(12):e51942.

113. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 2008;452(7187):564-70.

114. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313(5795):1929-35.

115. Peck D, Crawford ED, Ross KN, Stegmaier K, Golub TR, Lamb J. A method for high-throughput gene expression signature analysis. Genome Biol 2006;7(7):R61.

116. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res 2014;42(Web Server issue):W449-60.

117. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 2014;32(4):347-55.

118. Wright AV, Nunez JK, Doudna JA. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell 2016;164(1-2):29-44.

119. Xue W, Chen S, Yin H, Tammela T, Papagiannakopoulos T, Joshi NS, et al. CRISPR-mediated direct mutation of cancer genes in the mouse liver. Nature 2014;514(7522):380-4.

120. Platt RJ, Chen S, Zhou Y, Yim MJ, Swiech L, Kempton HR, et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell 2014;159(2):440-55.

121. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 2014;343(6166):84-7.

122. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science 2014;343(6166):80-4.

123. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, et al. Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis. Cell 2015;160(6):1246-60.

124. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat Biotechnol 2016;34(3):339-44.

125. Maruyama T, Dougan SK, Truttmann MC, Bilate AM, Ingram JR, Ploegh HL. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat Biotechnol 2015;33(5):538-42.

126. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol 2015;33(5):543-8.

127. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. N Engl J Med 2010;362(25):2380-8.

128. Van Allen EM, Wagle N, Levy MA. Clinical analysis and interpretation of cancer genome data. J Clin Oncol 2013;31(15):1825-33.

129. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304(5676):1497-500.

130. Douillard JY, Oliner KS, Siena S, Tabernero J, Burkes R, Barugel M, et al. Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. N Engl J Med 2013;369(11):1023-34.

131. Garraway LA. Genomics-Driven Oncology: Framework for an Emerging Paradigm. J Clin Oncol 2013.

132. Domchek SM, Greenberg RA. Breast cancer gene variants: separating the harmful from the harmless. J Clin Invest 2009;119(10):2895-7.

133. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17(5):405-24.

134. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499(7457):214-8.

135. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505(7484):495-501.

136. Hahn WC, Counter CM, Lundberg AS, Beijersbergen RL, Brooks MW, Weinberg RA. Creation of human tumour cells with defined genetic elements. Nature 1999;400(6743):464-8.

137. Dunn GP, Cheung HW, Agarwalla PK, Thomas S, Zektser Y, Karst AM, et al. In vivo multiplexed interrogation of amplified genes identifies GAB2 as an ovarian cancer oncogene. Proceedings of the National Academy of Sciences of the United States of America 2014.

138. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature 2013;503(7477):548-51.

139. Kang S, Denley A, Vanhaesebroeck B, Vogt PK. Oncogenic transformation induced by the p110beta, -gamma, and -delta isoforms of class I phosphoinositide 3-kinase. Proc Natl Acad Sci U S A 2006;103(5):1289-94.

140. Robles-Espinoza CD, Harland M, Ramsay AJ, Aoude LG, Quesada V, Ding Z, et al. POT1 loss-of-function variants predispose to familial melanoma. Nat Genet 2014;46(5):478-81.

141. Shi J, Yang XR, Ballew B, Rotunno M, Calista D, Fargnoli MC, et al. Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. Nat Genet 2014;46(5):482-6.

142. Ramsay AJ, Quesada V, Foronda M, Conde L, Martinez-Trillos A, Villamor N, et al. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. Nat Genet 2013;45(5):526-30.

143. Bainbridge MN, Armstrong GN, Gramatges MM, Bertuch AA, Jhangiani SN, Doddapaneni H, et al. Germline mutations in shelterin complex genes are associated with familial glioma. J Natl Cancer Inst 2015;107(1):384.

144. Calvete O, Martinez P, Garcia-Pavia P, Benitez-Buelga C, Paumard-Hernandez B, Fernandez V, et al. A mutation in the POT1 gene is responsible for cardiac angiosarcoma in TP53-negative Li-Fraumeni-like families. Nat Commun 2015;6:8383.

145. Greulich H, Kaplan B, Mertins P, Chen TH, Tanaka KE, Yun CH, et al. Functional analysis of receptor tyrosine kinase mutations in lung cancer identifies oncogenic extracellular domain mutations of ERBB2. Proc Natl Acad Sci U S A 2012;109(36):14476-81.

146. Stephen AG, Esposito D, Bagni RK, McCormick F. Dragging ras back in the ring. Cancer Cell 2014;25(3):272-81.

147. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative clinical genomics of advanced prostate cancer. Cell 2015;161(5):1215-28.

148. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A 2015;112(40):E5486-95.

149. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20(1):110-21.

150. Suzuki Y, Kagawa N, Fujino T, Sumiya T, Andoh T, Ishikawa K, et al. A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. Nucleic Acids Res 2005;33(12):e109.

151. Ward PS, Patel J, Wise DR, Abdel-Wahab O, Bennett BD, Coller HA, et al. The Common Feature of Leukemia-Associated IDH1 and IDH2 Mutations Is a Neomorphic Enzyme Activity Converting alpha-Ketoglutarate to 2-Hydroxyglutarate. Cancer Cell 2010;17(3):225-34.

152. Salmena L, Carracedo A, Pandolfi PP. Tenets of PTEN tumor suppression. Cell 2008;133(3):403-14.

153. Tilot AK, Gaugler MK, Yu Q, Romigh T, Yu W, Miller RH, et al. Germline disruption of Pten localization causes enhanced sex-dependent social motivation and increased glial production. Hum Mol Genet 2014;23(12):3212-27.

154. Xu J, Li Z, Wang J, Chen H, Fang J-Y. Combined PTEN Mutation and Protein Expression Associate with Overall and Disease-Free Survival of Glioblastoma Patients. Translational Oncology 2014;7(2):196-205.e1.

155. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 2012;44(6):685-9.

156. Le Gallo M, O'Hara AJ, Rudd ML, Urick ME, Hansen NF, O'Neil NJ, et al. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat Genet 2012;44(12):1310-5.

157. An J, Ren S, Murphy SJ, Dalangood S, Chang C, Pang X, et al. Truncated ERG Oncoproteins from TMPRSS2-ERG Fusions Are Resistant to SPOP-Mediated Proteasome Degradation. Mol Cell 2015;59(6):904-16.

158. Gan W, Dai X, Lunardi A, Li Z, Inuzuka H, Liu P, et al. SPOP Promotes Ubiquitination and Degradation of the ERG Oncoprotein to Suppress Prostate Cancer Progression. Mol Cell 2015;59(6):917-30.

159. Cancer Genome Atlas Research N, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature 2013;497(7447):67-73.

160. Zhang P, Gao K, Jin X, Ma J, Peng J, Wumaier R, et al. Endometrial cancer-associated mutants of SPOP are defective in regulating estrogen receptor-alpha protein turnover. Cell Death Dis 2015;6:e1687.

161. Geng C, Rajapakshe K, Shah SS, Shou J, Eedunuri VK, Foley C, et al. Androgen receptor is the key transcriptional mediator of the tumor suppressor SPOP in prostate cancer. Cancer Res 2014;74(19):5631-43.

162. Dang CV. MYC on the path to cancer. Cell 2012;149(1):22-35.

163. Delmore JE, Issa GC, Lemieux ME, Rahl PB, Shi J, Jacobs HM, et al. BET bromodomain inhibition as a therapeutic strategy to target c-Myc. Cell 2011;146(6):904-17.

164. Davis RJ, Welcker M, Clurman BE. Tumor suppression by the Fbw7 ubiquitin ligase: mechanisms and opportunities. Cancer Cell 2014;26(4):455-64.

165. Akhoondi S, Sun D, von der Lehr N, Apostolidou S, Klotz K, Maljukova A, et al. FBXW7/hCDC4 is a general tumor suppressor in human cancer. Cancer Res 2007;67(19):9006-12.

166. O'Neil J, Grim J, Strack P, Rao S, Tibbitts D, Winter C, et al. FBW7 mutations in leukemic cells mediate NOTCH pathway activation and resistance to gamma-secretase inhibitors. J Exp Med 2007;204(8):1813-24.

167. R Core Team. R: A language and environment for statistical computing: R Foundation for Statistical Computing; 2014.

168. Oliveros JC. Venny. An interactive tool for comparing lists with Venn's diagrams. 2007-2015.

169. Bowtell DD. The genesis and evolution of high-grade serous ovarian cancer. Nat Rev Cancer 2010;10(11):803-8.

170. Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, Dao F, et al. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474(7353):609-15.

171. Cheung HW, Cowley GS, Weir BA, Boehm JS, Rusin S, Scott JA, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. P Natl Acad Sci USA 2011;108(30):12372-77.

172. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang XP, et al. Highly parallel identification of essential genes in cancer cells. P Natl Acad Sci USA 2008;105(51):20380-85.

173. Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. Nature communications 2013;4:2126.

174. Anglesio MS, Wiegand KC, Melnyk N, Chow C, Salamanca C, Prentice LM, et al. Type-specific cell line models for type-specific ovarian cancer research. PLoS One 2013;8(9):e72162.

175. Etemadmoghadam D, Weir BA, Au-Yeung G, Alsop K, Mitchell G, George J, et al. Synthetic lethality between CCNE1 amplification and loss of BRCA1. Proc Natl Acad Sci U S A 2013;110(48):19489-94.

176. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483(7391):603-7.

177. Ren Y, Cheung HW, von Maltzhan G, Agrawal A, Cowley GS, Weir BA, et al. Targeted tumor-penetrating siRNA nanocomplexes for credentialing the ovarian cancer oncogene ID4. Sci Transl Med 2012;4(147):147ra12.

178. Karst AM, Levanon K, Drapkin R. Modeling high-grade serous ovarian carcinogenesis from the fallopian tube. Proceedings of the National Academy of Sciences of the United States of America 2011;108(18):7547-52.

179. Levanon K, Ng V, Piao HY, Zhang Y, Chang MC, Roh MH, et al. Primary ex vivo cultures of human fallopian tube epithelium as a model for serous ovarian carcinogenesis. Oncogene 2010;29(8):1103-13.

180. Flesken-Nikitin A, Hwang CI, Cheng CY, Michurina TV, Enikolopov G, Nikitin AY. Ovarian surface epithelium at the junction area contains a cancer-prone stem cell niche. Nature 2013;495(7440):241-5.

181. Hadari YR, Gotoh N, Kouhara H, Lax I, Schlessinger J. Critical role for the docking-protein FRS2 alpha in FGF receptor-mediated signal transduction pathways. P Natl Acad Sci USA 2001;98(15):8578-83.

182. Ong SH, Hadari YR, Gotoh N, Guy GR, Schlessinger J, Lax I. Stimulation of phosphatidylinositol 3-kinase by fibroblast growth factor receptors is mediated by coordinated recruitment of multiple docking proteins. P Natl Acad Sci USA 2001;98(11):6074-79.

183. Turner N, Grose R. Fibroblast growth factor signalling: from development to cancer. Nat Rev Cancer 2010;10(2):116-29.

184. Zhang KQ, Chu KV, Wu XW, Gao HL, Wang JH, Yuan YC, et al. Amplification of FRS2 and Activation of FGFR/FRS2 Signaling Pathway in High-Grade Liposarcoma. Cancer Res 2013;74(4):1298-307.

185. Wang XK, Asmann YW, Erickson-Johnson MR, Oliveira JL, Zhang HY, Moura RD, et al. High-Resolution Genomic Mapping Reveals Consistent Amplification of the Fibroblast Growth Factor Receptor Substrate 2 Gene in Well-Differentiated and Dedifferentiated Liposarcoma. Gene Chromosome Canc 2011;50(11):849-58.

186. Cappellen D, De Oliveira C, Ricol D, de Medina SGD, Bourdin J, Sastre-Garau X, et al. Frequent activating mutations of FGFR3 in human bladder and cervix carcinomas. Nat Genet 1999;23(1):18-20.

187. Kunii K, Davis L, Gorenstein J, Hatch H, Yashiro M, Di Bacco A, et al. FGFR2-amplified gastric cancer cell lines require FGFR2 and Erbb3 signaling for growth and survival. Cancer Res 2008;68(7):2340-48.

188. Dutt A, Salvesen HB, Chent TH, Ramos AH, Onofrio RC, Hatton C, et al. Drug-sensitive FGFR2 mutations in endometrial carcinoma. P Natl Acad Sci USA 2008;105(25):8713-17.

189. Weiss J, Sos L, Seidel D, Peifer M, Zander T, Heuckmann JM, et al. Frequent and Focal FGFR1 Amplification Associates with Therapeutically Tractable FGFR1 Dependency in Squamous Cell Lung Cancer (vol 3, 66er5, 2011). Science translational medicine 2010;2(62).

190. Dutt A, Ramos AH, Hammerman PS, Mermel C, Cho J, Sharifnia T, et al. Inhibitor-Sensitive FGFR1 Amplification in Human Non-Small Cell Lung Cancer. Plos One 2011;6(6).

191. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447(7148):1087-U7.

192. Sawey ET, Chanrion M, Cai CL, Wu GM, Zhang JP, Zender L, et al. Identification of a Therapeutic Strategy Targeting Amplified FGF19 in Liver Cancer by Oncogenomic Screening. Cancer Cell 2011;19(3):347-58.

193. Harding TC, Long L, Palencia S, Zhang HB, Sadra A, Hestir K, et al. Blockade of Nonhormonal Fibroblast Growth Factors by FP-1039 Inhibits Growth of Multiple Types of Cancer. Science translational medicine 2013;5(178).

194. Hagerstrand D, Tong A, Schumacher SE, Ilic N, Shen RR, Cheung HW, et al. Systematic interrogation of 3q26 identifies TLOC1 and SKIL as cancer drivers. Cancer discovery 2013;3(9):1044-57.

195. Lowenstein EJ, Daly RJ, Batzer AG, Li W, Margolis B, Lammers R, et al. The Sh2 and Sh3 Domain Containing Protein Grb2 Links Receptor Tyrosine Kinases to Ras Signaling. Cell 1992;70(3):431-42.

196. Warner N, Nunez G. MyD88: A Critical Adaptor Protein in Innate Immunity Signal Transduction. J Immunol 2013;190(1):3-4.

197. Kouhara H, Hadari YR, SpivakKroizman T, Schilling J, BarSagi D, Lax I, et al. A lipid-anchored Grb2-binding protein that links FGF-receptor activation to the Ras/MAPK signaling pathway. Cell 1997;89(5):693-702.

198. Cheung HW, Du JY, Boehm JS, He F, Weir BA, Wang XX, et al. Amplification of CRKL Induces Transformation and Epidermal Growth Factor Receptor Inhibitor Resistance in Human Non-Small Cell Lung Cancers. Cancer discovery 2011;1(7):608-25.

199. Dunn GP, Cheung HW, Agarwalla PK, Thomas S, Zektser Y, Karst AM, et al. In vivo multiplexed interrogation of amplified genes identifies GAB2 as an ovarian cancer oncogene. Proc Natl Acad Sci U S A 2014;111(3):1102-7.

200. Chen Y, McGee J, Chen X, Doman TN, Gong X, Zhang Y, et al. Identification of Druggable Cancer Driver Genes Amplified across TCGA Datasets. Plos One 2014;9(5):e98293.

201. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6(269):pl1.

202. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery 2012;2(5):401-4.

203. Liu JS, Yang G, Thompson-Lanza JA, Glassman A, Hayes K, Patterson A, et al. A genetically defined model for human ovarian cancer. Cancer Res 2004;64(5):1655-63.

204. Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, et al. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. Genomics 2007;89(3):307-15.

205. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology 2006;7(10):R100.